DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Essays on Causality, Race, and the Law

Citation
Sen, Maya. 2012. Essays on Causality, Race, and the Law. Doctoral dissertation, Harvard University.

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:9306421

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

©2012 – Maya Sen
All rights reserved.

Essays on Causality, Race, and the Law

Abstract

Making causal inferences about race is difficult because no means exist to manipulate units into treatment and control groups. Chapter 1 addresses this predicament. First, I argue that race should be defined as a composite measure in which some elements are mutable. Second, I note that identifying the units of analysis is particularly important when thinking about race. These extensions allow us to synthesize two instances in which causal inferences regarding race may be permissible: (1) studies that measure the effect of exposing an entity to a racial cue and (2) studies that disaggregate race into constituent pieces and measure the effect of a mutable element.

Chapters 2 and 3 provide examples of the first "exposure" approach in the context of judicial politics. Chapter 2 analyzes the role of race and gender in judicial confirmations and demonstrates that minority and female nominees to federal courts are awarded lower qualification ratings by the American Bar Association (ABA) than are white and male nominees. This is the case even when comparing only judges with similar education, ideologies, and experiences. Furthermore, I present results showing that ABA qualification scores are not predictive of judges' reversal rates.

Chapter 3 explores what happens once minority judges are confirmed. Focusing specifically on African Americans, I show that opinions authored by black judges are overturned more than cases authored by whites. The effect is robust and persists after taking into account measures of judicial qualifications, previous professional and judicial experience, and partisanship. Taken together, Chapters 2 and 3 have clear implications: despite attempts to make judiciary more reflective of the U.S. population, racial disparities continue to persist.

Contents

Author List

The following authors contributed to Chapter 1: Maya Sen and Omar Wasow
The following authors contributed to Chapter 2: Maya Sen
The following authors contributed to Chapter 3: Maya Sen

Listing of figures

To my family and friends.

Acknowledgments

I am very grateful to the members of my committee: Jennifer Hochschild, Gary King, Kevin Quinn, and Adam Glynn. In addition, I have had numerous helpful conversations about this dissertation project with Matt Blackwell, Amy Catalinac, Andrew Coe, Porsha Cropper, Ryan Enos, Bernard Fraga, Justin Grimmer, Nahomi Ichino, Konstantin Kashin, Jenn Larson, Ryan T. Moore, Clayton Nall, Rich Nielsen, Ellie Powell, Jonathan Renshon, Shauna Shames, Arthur Spirling, Dustin Tingley, Brandon Van Dyck, and Miya Woolfalk. I am especially grateful to my colleague Omar Wasow, with whom portions of the first chapter were co-authored.

This project has also benefitted from presentations at the Harvard American Politics Research Workshop, the Harvard Applied Statistics Workshop, the Midwest Political Science Association annual conference, the Society for Political Methodology summer meeting, and the Conference on Empirical Legal Studies. I am also deeply appreciative for the hundreds of questions, suggestions, and criticisms that I received at presentations at UC-Berkeley, UCLA, George Washington, the University of Iowa, Ohio State, Rice, Vanderbilt, UVA, the University of Rochester, MIT, Stanford, Princeton, NYU, and the University of Colorado. This project is stronger because of that feedback.

I am also incredibly grateful for financial support over the years from the John M. Olin Center for Law, Economics, and Business at Harvard Law School, the Robert Wood Johnson Foundation, the Institute for Quantitative Social Science, the Harvard Center for American Political Studies, and the Harvard Graduate School of Arts and Sciences.

Lastly, I extend a warm thank you to my family and friends.

1

How to Extract Causal Inferences About Race

Experts in causality have long warned against making causal inferences on the basis of race or ethnicity. Why? Race is commonly understood as an immutable or unchanging characteristic. For centuries, societies have categorized people from birth as belonging to different groups like black, white, Native American, Aboriginal, etc., and a person's race is generally thought to be resistant to manipulation. As a result, it is impossible to randomly assign some people to one race and others to a different race. From a causal perspective, this means it is infeasible to assess how someone would fare

if black as opposed to white, Eskimo as opposed to Australian Aboriginal, or Asian as opposed to Native American without taking into account the myriad cultural, educational, political and economic differences between people of different backgrounds. Furthermore, because factors like education and class are intimately linked to the distinct historical experiences of each racial and ethnic group, the whole endeavor can become a tangled mess – at least when attempting to make causal claims. The end result, as experts have long warned, is that making causal inferences about race and race-based variables is exceedingly difficult, if not impossible.

For many applied scholars, however, making causal inferences about the role of race and other immutable characteristics lies at the core of key research questions. Researchers from fields as disparate as comparative politics, race and ethnic politics, economics, psychology, sociology, public policy, law, and health policy routinely focus their attention on the causal role(s) of race. For researchers exploring these sorts of questions, the primary inquiry lies precisely in teasing apart the causal effects of race or ethnicity on a wide variety of outcomes.

Persistent interest in estimating the causal effects of immutable characteristics leaves methodologists and statisticians stuck between a rock and a hard place. Methodologists can either continue hand-wringing about the naïve (and maybe biased and misleading) causal claims made by applied researchers. Or, as a second possibility, methodologists can analogize to successful experimental studies and move forward by setting standards under which permissible – and perhaps quite limited – causal inferences can be made.

We choose the second option, and our hope is that this paper begins to bridge the divide between cautious causal experts on the one hand and, on the other, researchers for whom causal questions involving immutable characteristics such as race are simply too important to set aside. To that end, while we are largely in agreement with the

extant warnings about using race as a treatment variable, we argue that existing thinking about race and causality has overlooked two key points. First, applied researchers have under-theorized the appropriate unit of analysis and, as a result, may be overlooking particularly appropriate experimental analogies. Second, empirical work addressing race tends to do so with little attention to how race is defined. Within quantitative social science, race is usually a single monolithic variable – similar to GDP per capita, or the dosage in a vaccine, or attending a job training program. But as scholars from a wide range of fields have noted (e.g., Appiah (1986); López (1994); Holland (2003)), racial and ethnic categories are typically the product of a complicated amalgam of social, cultural, historical, biological, geographical and legal influences. Speaking metaphorically, we think that race can be viewed as a "bundle of sticks" instead of a single bolt of wood. Rather than gloss over the complex challenges of measuring race (as frequently occurs in quantitative scholarship), we argue in favor of exploiting its composite nature by disaggregating race into constituent elements that can be reasonably defined and manipulated.

Thinking more flexibly about these two considerations opens up the potential outcomes framework to handle a variety of instances in which causal questions involve characteristics like race or gender. To be clear, this is it not to say that making causal inferences about race is possible in all instances; what we argue in this paper that there are some limited areas in which making causal inferences is permissible *provided that the researcher has given careful thought to (1) potential post-treatment problems, (2) the appropriate unit of analysis, (3) the composition of the treated and control groups, and, perhaps most importantly, (4) what he or she means by "race."*

This paper proceeds as follows. First, we explain briefly the potential outcomes framework and note the intrinsic problems involved with making causal inferences

about race or gender. Second, we move forward by noting that race brings with it different ways of thinking about experimental units, treatment regimes, and the composition of appropriate treated and control populations. Tying these threads together, we then develop a framework that unifies work from a wide range of disciplines into two types of studies that estimate effects of race: (1) studies that measure the effect of exposing an individual or institution to some signal about race and (2) studies that disaggregate race into constituent pieces and attempt to measure the effect of some mutable element of race within a single racial group. Additionally, by analogizing to effective experimental designs, we clarify how race-based variables can – and cannot – be used in extracting causal inferences from observational studies. The later sections of the paper develop these ideas through empirical examples. We conclude by outlining areas of future research.

1.1 Causal Inferences and Potential Outcomes

A brief overview[1] of the potential outcomes framework helps contextualize the following discussion. At its core, a causal inquiry involves unpacking the effect of some treatment on some outcome. Does a vaccine cause people to live longer? Will a drug treatment program prevent addicts from using drugs or alcohol? Is a worker training program effective in helping people go back to work? In all of these cases we see (1) a unit of analysis, (2) a manipulable treatment, and (3) a specific outcome.

1.1.1 The fundamental problem of causal inference

Some basic notation illustrates these notions. For a unit i, we are ultimately interested in the effect of some treatment, T_i, on an outcome, Y_i. The treatment may be binary –

[1] The literature on this topic is voluminous – e.g., Rubin (1974), Holland (1986), Angrist, Imbens and Rubin (1996), Rubin (2005), and we attempt in this paper only a brief summary.

e.g., worker i is offered a spot in the worker training program or she is not, subject i is injected with a vaccine or he is not – resulting in $T_i = \{1, 0\}$. Assuming a binary treatment variable, a unit i can either be treated $(T_i = 1)$ or not treated $(T_i = 0)$, giving us two "potential outcomes," $Y_i(T_i = 1)$ and $Y_i(T_i = 0)$.

In an ideal world, we would want to estimate the *true* treatment effect, or the simple difference between these two potential outcomes for unit i as specified in Equation 1.1:

$$Y_i(T_i = 1) - Y_i(T_i = 0) \tag{1.1}$$

The "fundamental" problem of causal inference is, however, that we can never observe the difference between $Y_i(T_i = 1)$ and $Y_i(T_i = 0)$ (Holland, 1986; Rubin, 1978). That is, unit i cannot receive both the treatment and the control – i.e., a study participant cannot be given the vaccine *and* also be given a placebo, and a worker cannot both be offered a spot in a worker training program *and* be rejected.

We note here that the fundamental problem of causal inference extends to all kinds of inquiries – for example, when testing different medicines, or seeing the effects of a work training program – but it becomes particularly vexing when it comes to immutable characteristics. After all, a person cannot experience the world as being only black and also as being only white, or as being only Native American and as being only Maori, and to think otherwise raises strange hypotheticals. On this point, Holland (2003) raises, and quickly dismisses, the potential lessons learned from ostensibly causal narratives such as *Black Like Me* and *Soul Sister*.[2] This is an important point to

[2] We may think that the experiences of mixed-race people may solve the fundamental problem of causal inference. After all, people who are mixed race routinely "pass" as members of one group and then also as members of another group, and a rich and varied literature (scholarly as well as popular) has developed around how multiracial people self-identify (Faulkner, 1990; Schuyler, 1971; Halsell, 1969; Griffin, 1996; Gates, 1997; Kim and Lee, 2001; Hochschild and Weaver, 2010). From a causal perspective, however, mixed race people represent those for whom a third kind of "treatment" has been administered – a mixed race treatment. Thus, although the experiences of these sorts of individuals may be informative – and illus-

which we return again later on in the discussion.

In lieu of trying to estimate an unobservable true treatment effect, those interested in causal estimates have moved forward by estimating some version of the *average* treatment effect,

$$E[Y_i(T_i = 1)] - E[Y_i(T_i = 0)], \tag{1.2}$$

that is, Equation 1.2 calculates the difference between the outcome means in treated and control populations. (Variants of the average treatment effect include average treatment effect on the treated (ATT), local average treatment effects (LATE), and sample average treatment effects (SATE).) An obvious problem is, however, that differences in the outcome variable could be due to inherent differences between the treated and control populations. For example, we should not be surprised to see that workers who have signed up for a worker training program are more successful in getting jobs – but we also should not be surprised that individuals who have enrolled in worker training programs are both more ambitious and better educated than non-trained workers, two attributes that would also result in more favorable employment decisions.

The problem is solved in some circumstances by comparing only similarly situated treated and control units. Let X_i represent the background variables that could affect both the probability of receiving treatment or the eventual outcome. To get at a satisfactory estimate of the average treatment effect, we would like our treatment and control groups to be so similar across X such that that the only difference between the two groups is that one received the treatment and the other did not. This ignorability

trates that the neat disaggregation of races into categories is never easy, a point we develop below – much of this discussion applies to multiracial individuals just as it does people who self-identify exclusively as black, white, Asian American, etc.

assumption is stated in Equation 1.3 as:

$$P(Y_i(T_i = 1), Y_i(T_i = 0)|X_i, T) = P(Y_i(T_i = 1), Y_i(T_i = 0)|X_i), \tag{1.3}$$

i.e., if $Y_i(T_i = 1)$ and $Y_i(T_i = 0)$ are independent of T_i, conditional on the covariates X_i (Holland, 1986). In plain English, the treatment assignment must be independent of the potential outcomes in order for us to assume that the two groups are similar enough to extract causal inferences.

The challenge for social science researchers working with observational data is usually to satisfy the ignorability requirement – that is, to make the treated and control populations as similar as possible so that the treatment regime could be assumed to be random. By far the easiest way to satisfy the ignorability requirement is simply to assign the treatment randomly – for example, by conducting a randomized experiment. (We discuss some successful experimental designs below; more general discussions are found in Holland (1986) and Imai, King and Stuart (2008).) Using a job training program example, a researcher could flip a coin and randomize some workers to be accepted into a worker training program and for others to be rejected. Such an approach would ensure that the subset of workers accepted into the training program would be identical to the group of rejected workers across all possible observed and unobserved variables. Because randomization is often times not an option for social scientists, researchers have turned to matching on observed variables (using propensity score, genetic, nearest neighbor, exact, or coarsened matching) to satisfy the ignorability assumption that the treated and control groups are identical on background covariates[3] (Dehejia and Wahba, 2002; Sekhon, 2009).

[3] Matching, as others have noted, also has the benefit of reducing model dependence in parametric models (Ho et al., 2007).

1.2 Race and Potential Outcomes

Prior work on race and causation has identified two key problems that arise when using race as a treatment within the potential outcomes framework (Greiner and Rubin, 2010). First, biological elements of race, such as skin color or facial features, are resistant to manipulation. This fact makes it difficult to conceptualize well-defined potential outcomes.[4] Second, because much of race is generally understood to be "assigned" at birth (or conception), the host of characteristics for which most social scientists control (e.g., education, income, etc.) occur after the treatment is assigned and therefore have the potential to introduce post-treatment bias.

We propose a third concern with race and causation. Building on the idea that race should be viewed as an amalgam of characteristics or a "bundle of sticks," we argue that an additional difficulty in imagining race as a treatment is that researchers often misclassify what the race variable represents. We discuss all three problems below.

1.2.1 Conceptualizing the appropriate potential outcomes

The potential outcomes framework – and making causal inferences in general – demands a neatly defined, manipulable treatment variable, one that can be easily documented and manipulated by researchers.[5] Holland (1986), for example, famously admonishes "No causation without manipulation" to bring attention to the idea that all

[4] A large literature in gender studies distinguishes between "sex" and "gender" where "sex" is defined as biological and anatomical while "gender" is defined as the product of psychological, social and cultural forces – see, for example, (West and Zimmerman, 1987; Deaux, 1985). We similarly distinguish between more immutable, biological elements of race and more mutable, socially defined elements of race.

[5] The literature on this point is rooted as much in statistics as it is in philosophy and political thought – e.g., Locke (1847), Hume (2003), and Mill (1884). More recently, philosophers looking at the topic have advanced the idea that manipulation is at the core of a causal inquiry, and treatments such as Menzies and Price (1993) and Von Wright (1971) focus specifically on human intervention or action. Others, such as Hausman (1998), have critiqued this literature as putting too much emphasis on human agency, which has the effect of overlooking the natural or non-human manipulations and interventions that can occur. Holland (1986) provides an informative overview of this literature.

pertinent potential outcomes must be defined in principle in order to make causal estimates possible in practice. Further, to define all potential outcomes, one must be able to conceptualize an experimental analogy that would lead to the possible outcomes. In other words, as Holland puts it, "causes are only those things that could, in principle, be treatments in experiments." This idea of a manipulative treatment is echoed by others like Cook, Campbell and Day (1979), who argue that "[c]ausation implies that by varying one factor I can make another vary"; Pearl (2000), which discusses at length the importance of an intervention in estimating causal treatments; and Gelman and Hill (2007), who warn that "a causal effect needs to be defined with respect to a cause, or an intervention, on a particular set of experimental units."

The biological dimensions of race and gender are, however, resistant to manipulation.[6] Treatment by race and gender also suffer from the problem that it is difficult to think about appropriate counterfactuals. We can imagine how someone lives their life as an African American; much more difficult is imagining what experiment one would design to manipulate the person's race (and *only the person's race*) to check its effect on some outcome. Thus, not only is randomization, the most elegant solution to the fundamental problem of causal inference, beyond the reach of those scholars focusing on the effects of race or ethnicity, but the very idea of thinking about causality and race is, at its core, a problematic enterprise.

The immutable (i.e., resistant to manipulation) nature of race and gender has led those demanding manipulation to cite race and gender as attributes for which causal inferences are impermissible (e.g., Holland (1986); Rubin (1978); Gelman and Hill (2007)).[7] As noted by Holland (1986): "For causal inference, it is critical that each

[6] In an unpublished manuscript, Imbens and Rubin (2010) refer to "currently immutable characteristics" as future innovations may dramatically ease the effort required to change to certain biological aspects of race or gender.

[7] We note, however, that these warnings do not apply to predictive or correlative inferences, and both

unit be potentially exposable to any of the causes. As an example, the schooling a student receives can be a cause, in our sense, of the student's performance on a test, whereas the student's race or gender cannot." A more specific admonishment on the topic of gender-based causality is given by Rubin (1978):

> [C]onsider the causal effect of sex (male-female) on intelligence. What are the actions to be applied to each experimental unit that define the treatments? Are we to give hormone shots beginning at birth and surgically perform a "sex-change" operation, or at conception "change" Y-chromosomes and X-chromosomes? Even if an "at-conception X- for-Y chromosome change" becomes possible, presumably there will be several techniques developed for effecting the change with potentially different causal effects. Without treatment definitions that specify actions to be performed on experimental units, we cannot unambiguously discuss causal effects of treatments.

Thus, the difficulty of conceptualizing well-defined potential outcomes means that many are fundamentally skeptical about making causal inferences about race and other seemingly immutable characteristics.

1.2.2 Problems with post-treatment bias

The Rubin/Holland objection to attributes-based causal inference has received some pushback (Heckman, 2005; Greiner and Rubin, 2010) and we discuss some of these approaches in greater depth below. In addition, some applied researchers have eschewed warnings about the immutable attributes (and the accordant problems with conceptualizing possible counterfactuals) by using matching on background covariates – e.g., Boyd, Epstein and Martin (2010). There is, however, a problem secondary to conceptualizing well-defined potential outcomes: a person's race is "assigned" by a combination of social and biological processes at conception or birth. Thus, the host of

Rubin and Holland would likely agree that interesting questions can be asked and explored using these non-causal techniques.

background covariates that social scientists usually control for and match on (e.g., education, income, age) are determined *after* a person's race is assigned.

Taking into account things that happen after the treatment happens or is administered raises the specter of post-treatment bias, a pervasive problem in the social sciences (King, Keohane and Verba, 1994; Rosenbaum, 2002). To use a common example, suppose that we are interested in the causal effect of smoking on death, and have a population of randomly assigned smokers and randomly assigned non-smokers. Would we want to control for lung cancer in the final analysis? Probably not: lung cancer is not only highly predictive of death, but it is also a direct consequence of smoking – probably *the* key consequence. If we controlled for lung cancer, then the effect of smoking on death would essentially be nill, biased downward by the fact that we have controlled for its primary consequence. Such post-treatment bias arises when we control for things (like lung cancer) that are a direct consequence of the treatment. Race is obviously different from smoking, but the post-treatment issue applies with equal or greater force: race affects deeply how a person is raised and educated, what kinds of employment opportunities (and hence employment experiences) he or she will have, and what kind of cultural and social attitudes he or she will bring to the table. Including any of these attributes would therefore affect our estimates of the causal effect of "race."[8]

[8] Even aside from the post-treatment issue, two related problems with this kind of strategy are (1) common support problems and (2) problems with multicollinearity. The common support problem arises when researchers include attributes that vary according to race (e.g., welfare status, participation in programs like Head Start, diseases such as sickle cell anemia or Tay Sachs). Because these traits vary almost exactly according to race, it becomes difficult to find non-minority (or minority) counterparts with which to compare the population of interest. (For example, it would be hard to find a group of whites who have sickle cell anemia (Thomas and Zarda, 2011).) Collinearity becomes a problem when variables or effects vary so closely with race so as to result in (the most extreme case) unconverged calculations of point estimates. The lack of variance in the background variables may also result in small changes having a large impact on the coefficient estimates – thus, standard errors may be large and lead researchers to assume no treatment effects when treatment effects do in fact exist.

Although perhaps unsatisfcatory to many applied researchers, the most appropriate initial approach is to drop any post-treatment variables from an analysis (King, 1991; King, Keohane and Verba, 1994; King and Zeng, 2006; Gelman and Hill, 2007). (We discuss alternate approaches and provide more detailed examples below.) Thus, in the race context, any factor, attribute, personality trait, or personal or professional experience that could potentially be a consequence of race should be dropped – a requirement that includes most of the variables included by social scientists.[9]

Going back to the employment example, suppose a researcher is interested in the way that individual African Americans fare when looking for a job. A straightforward least squares or logit regression would position employment decisions as the outcome variable and, on the right-hand side, include a black-white (or black versus non-black) "dummy" variable. Also included would be a slew of additional "control" variables such as gender, highest education level, age, and job training experience. A simple regression would look something like Equation 1.4:

$$Employment_i = \alpha + \beta_1 * Race_i + \beta_2 * Gender_i + \beta_3 * Educ_i + \beta_4 * Age_i + \beta_5 * Training \quad (1.4)$$

where the coefficient on the race dummy, β_1 is the estimate of interest. Now, with post-treatment problems squarely in mind, the problem with this specification is that variables like education, age, and skills are post-treatment: being black will affect whether someone attains a certain level of education, how much he or she will make, what age he or she will live to, and what kind of job training programs will be available to him or her. Thus, the inclusion of these variables has the effect of biasing our estimate, possibly in a downward direction. A better specification would be one in

[9] This strategy implies that the researcher is interested in the *total* effect of race. There may be instances where this is not the case, and we discuss one of them in our section on exposure studies, below.

which these post-treatment variables are dropped, leaving only race and possibly gender[10] as the remaining explanatory variables:

$$Employment_i = \alpha + \beta_1 * Race_i + \beta_2 * Gender_i \tag{1.5}$$

We note that this approach does not counter fully the Rubin/Holland complaint, discussed above, that the potential outcomes are not well-defined. Nonetheless, dropping any post-treatment variables, as is done in Equation 1.5, is a useful first step for anyone trying to ascertain with rigor the total effect of race on an outcome variable. We also note that, in the case of race and other immutable characteristics, nearly all variables will be post-treatment, and should therefore be dropped – a harsh medicine, to be sure. The end result may be a somewhat unsatisfying analysis for many applied researchers and policy analysts, but it does represent the best way to estimate the total effects of race on some outcome variable.

1.2.3 Problems with race as a "bundle of sticks"

The prior two methodological problems associated with race are well-known in the causal inference literature. As noted earlier, we argue that race is generally not a single, easily defined measure but rather a composite of many component pieces or,

[10] Sex, which is also assigned at birth, is the only variable that could be possibly construed as being pre-treatment or, at the very least, assigned concurrently with the treatment. Evidence suggests, however, that sex ratios, or the ratio of boys to girls in a population, can vary by latitude, religion, ethnicity and other factors that may be collinear with race (Guttentag and Secord, 1983; Navara, 2009). It would be appropriate in instances such as this – where the timing or causal connection of the variable is somewhat in question – to try a specification that includes the variable and one that drops the variable, comparing and contrasting the results.

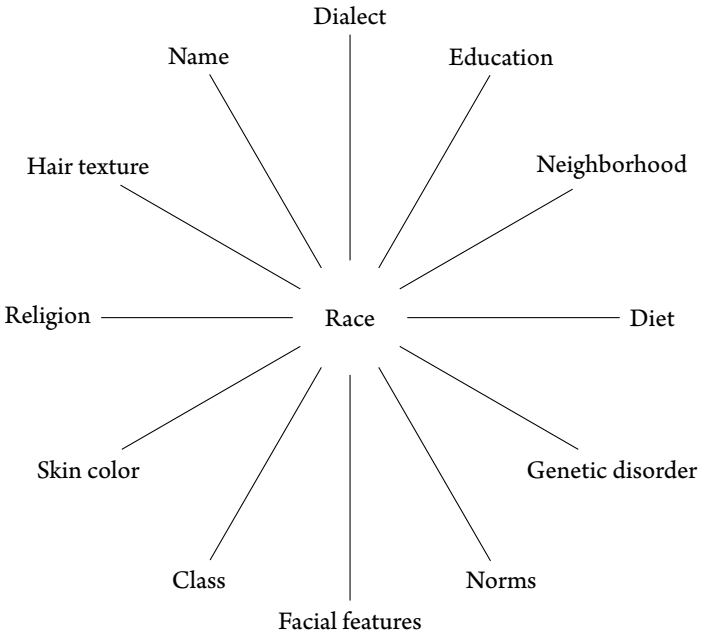

Figure 1.1: Some characteristics often associated with race or ethnicity, i.e., race as a "bundle of sticks."

metaphorically, a "bundle of sticks."[11] This composite conception poses a third problem when attempting to incorporate race as a treatment in the context of causal inference. Specifically, researchers could never assign the full bundle of factors that constitute a racial identity to some subjects while assigning others to a control. Indeed, changing an experimental unit's skin color, appearance, cultural background, language, expectations, attitudes and educational levels and so on is essentially impossible – let alone doing it in a way that approaches a truly random process. To illustrate the idea of race as a "bundle of sticks," Figure 1.1 presents a simple radial diagram of some characteristics that are constitutive of race.

Additionally, in much of quantitative social science, race variables are usually

[11] Legal scholars will no doubt understand this reference, as the bundle of sticks metaphor is frequently used to discuss property rights, which have multiple dimensions in different contexts.

represented by indicator dummy variables (e.g., "1" if black, "0" if white), categorical variables (e.g., "1" if white, "2" if black, "3" if Native American, etc.) or percentages (e.g. the percentage of the population in a given region that belongs to some racial group). The challenge of using such monolithic measures of race is that any statistical association will typically offer little or no insight as to which element of race is the key mechanism of action. For example, in the economic literature on crime, a standard practice is to include some right-hand side control variable of the "percent black" of the geographic units. Rarely, however, is any attempt made to interpret why such a measure is important or what particular forces might be driving the evident statistical relationships between race and crime. Quantifying race into binary, categorical or percentage variables (like, "Democrat" or "Republican"), as is routinely done in social science research, ignores the fact that racial identity is a multi-dimensional, variegated "variable" that means different things in different contexts.

1.3 Extending Potential Outcomes to Race

In the introduction, we argued that applied researchers attempting to make causal claims about "effects of race" have payed insufficient attention to defining race and to delineating the appropriate units of analysis. Building upon those comments, we now argue that, in a limited set of cases, careful attention to these two issues opens up the potential outcomes framework to causal inference with race. As we outline below, these extensions illuminate two possible ways to attack causal questions regarding race.

1.3.1 Exploiting the "bundle of sticks" with race and causation

Though defining race as a "bundle of sticks" may make causal inference harder, such a definition also has the potential to help reconcile causal inference with race in specific

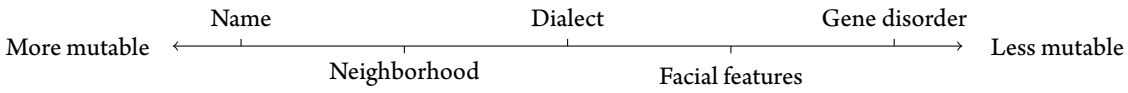

Figure 1.2: Hypothetical mutability of some characteristics associated with race and ethnicity.

cases. First, while it would be impossible to meaningfully assign all of the components of race as a treatment, disaggregating race allows for the investigation of an effect of a single "stick" or element of race. Second, disaggregating race reveals that some components of the "bundle" are more mutable than others which, in turn, enables manipulation. Figure 1.2 presents a hypothetical continuum of features that are strongly associated with race but that exhibit varying degrees of mutability. For example, facial features – such as the shape of one's eyes or the contours of one's nose – are fairly immutable, possibly changed through plastic surgery but certainly not something researchers could manipulate easily.[12] Usefully for purposes of causal inference, though, some of these elements of race, such as neighborhood, are mutable (Katz, Kling and Liebman, 2001), an aspect we discuss below.

Prior scholarship touching on race and causality has tended to view race largely in biological terms. Gelman and Hill (2007), for example, offer the case of height as an immutable characteristic. While height might be an appropriate analogy for biological features of race like skin color, it may be less appropriate for other aspects of race that are socially determined and, therefore, more mutable. Put simply, the analogy to height is inadequate when thinking about the whole "bundle of sticks."[13] Holland (2003), by

[12] Despite the obvious challenges, consumption of plastic surgery and other forms of race-bending body manipulation are increasingly popular and widespread and, therefore, may present growing opportunities for research designs that allow for causal inference. For example, a recent *New York Times* article suggests that a boom is underway in plastic surgery consumption among immigrant and ethnic communities in the United States (Dolnick, 2011). Similarly, according to a survey by Survey (2004), "38% of women surveyed in Hong Kong, Korea, Malaysia, the Philippines and Taiwan use some kind of skin lightening products."

[13] As the cliometrics literature has shown, height, itself, is also a composite measure of comprised of environmental factors like health and access to nutrition as well as biological factors like genetics (Fogel,

contrast, clearly recognizes the hybrid nature of race when noting, "I regard race as a socially determined construction with complex biological associations. I also believe that it is very naïve to disregard the durability and power of social constructions" Holland (2003, p. 3). We concur with Holland's assessment and argue that these durable and powerful social constructions are also often mutable features of race that have been under-exploited for causal inference.

To give an example, a person's name is relatively mutable but also typically provides a strong signal about racial or ethnic background (Chang et al., 2010). Although challenging, one could imagine an experiment in which new parents of the same race and background were randomly assigned to pick a baby name from one of two lists. One list would include names that are not strongly identified with the relevant racial or ethnic group and the other list would include names that do exhibit such an association. This kind of study could then assess the short- and long-run effects of a racially or ethnically specific name on outcomes like education or employment. In sum, not all of the "sticks" are inherently immutable; nor is the whole "bundle" automatically assigned at birth.

1.3.2 Delineating the appropriate unit of analysis

The other important consideration for reconciling causal inference and race is to clarify the unit of analysis and, by extension, the composition of the treated and control groups. Figure 1.3 presents a flowchart for considering two different units of analysis and whether the study is experimental or observational. In Figure 1.3, we suggest that at least two different research designs can successfully pose causal questions about an effect of race.

1994).

The first type of research design we call "exposure to race" or exposure studies. These studies examine how subjects respond when exposed to some sort of racial signal or cue and, as such, might be more precisely called "exposure to a racial signal" or "exposure to a racial cue" studies. Though exposure to race is a useful shorthand, it is important to note that the treatment is never race in full (i.e., the whole "bundle of sticks") but rather only an element of race. Among the studies included in this group would be those that look at how voters respond when presented with advertisements showing black versus white candidates, or those that examine whether employers are more or less likely to interview job applicants with traditionally African American names.

In the second type of study the appropriate unit of analysis is a single racial group that exhibits within-group variation on some element of race. This within-group variation can be achieved experimentally through treatment or identified in observational data. We refer to these types of studies as "element of race" studies because the treatment involves manipulating or observing variation in some constitutive element of race within a single racial group.

1.4 Treatment by Exposure to Race

By exposure study we mean research designs where exposure to a racial cue is the treatment of interest, and the unit of analysis is the individual or institution being exposed. Going back to our employment example, suppose we are ultimately interested in discrimination against black job candidates. A researcher conducting the ideal experiment would take two applicants, one white and one black, and construct a job profile that is exactly the same for each applicant – except for some signal or cue (such as a name or picture) that one applicant is black and one is white. (This is what

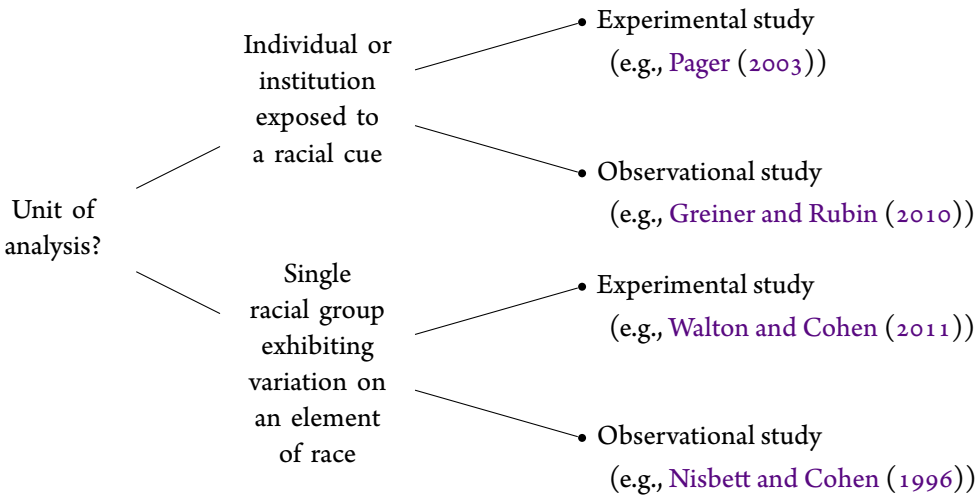

Figure 1.3: Thinking about the appropriate unit of analysis.

was done in Bertrand and Mullainathan (2004), which relied on distinctively African-American names to send a signal about the applicant's race.) The researcher would then send these job applications to employers and check the eventual hiring decisions. A difference would suggest that similarly situated blacks and whites are being treated differently, while no difference would suggest no discrimination. Key to this kind of study is that the unit of analysis is a prospective employer (*not* the prospective employ*ee*), and the treatment is the kind of name attached to the job application. Thus, the research design begins with well-defined potential outcomes and with a precise moment of treatment.

1.4.1 Experimental studies with treatment by exposure to race

These kinds of audit or correspondence studies have been used to measure race and gender discrimination in fields like employment (Neumark, Bank and Van Nort, 1996; Pager, 2003) and housing (Yinger, 1986). Although audit and correspondence studies

generally work with actors or fake applicants,[14] the principle of experimentally exposing a subject to a racial cue or signal is, however, more general, and it can be used in a wide variety of contexts. Indeed, exposing subjects to a racial cue or signal has been leveraged to estimate race-based causal effects across a variety of disciplines, including economics, (Bertrand and Mullainathan, 2004), sociology (Bobo and Johnson, 2004), psychology (Cosmides, Tooby and Kuzban, 2003; Boker et al., N.d.; Steele, 1997; Greenwald, McGhee and Schwartz, 1998).

Within political science, a robust public opinion literature (Miller and Krosnick, 2000; Brader, Valentino and Suhay, 2008; Huber and Lapinski, 2006; Valentino, Hutchings and White, 2002; Mendelberg, 2001; Sniderman and Piazza, 1993; White, 2007) has exploited the exposure research design to estimate race-based causal effects. Sniderman and Piazza (1993), for example, leverages the order in which questions about affirmative action and attitudes toward blacks are asked, finding that the "mere mention" of affirmative action to white survey respondents provokes more negative feelings towards blacks. Mendelberg (2001) creates simulated television news experiments to assess how racial cues might prime racial attitudes among white voters. Within psychology, Kurzban, Tooby and Cosmides (2001) expose subjects to photos and text to simulate a cross-race conversation and Steele (1997) identifies how internalized racial stereotypes affect women and racial minorites by exposing them to racial and gender cues immediately prior to a mathematics exam. Kurzban, Tooby and

[14] Pager (2007) provides a good overview of the literature, critiques, and methods. Although the exact methodology may vary, audit studies usually involve confederates or actors hired by researchers who are then randomly sent out to the field – for example, to different employers or to different lending agents. Partly in response to critiques about potential bias introduced by the confederates (Heckman and Siegelman, 1993; Heckman, 1998), correspondence studies were developed in which matched human applicants were replaced with matched pairs of "paper" applicants. Bertrand and Mullainathan (2004), as noted above, randomly assigned traditionally white and black names to otherwise similar resumes to assess how such signals about the race of the applicant affected hiring decisions. More recently, Adida, Laitin and Valfort (2010) used a similar technique to measure employment discrimination against Muslims in France.

Cosmides and Steele also demonstrate how the exposure model can address questions other than concerns about a discriminatory "decisionmaker" (to use the terminology of Greiner and Rubin (2010)).

To step back a moment, although all of these studies[15] exploit different techniques – from simulated avatars to scenarios in surveys – the general approach is the same: randomly present a subject with information that differs primarily with respect to signals or cues about race. Research designs of the exposure type thus have (1) a randomly assigned treatment, which is the racial signal or cue, and (2) a unit of analysis, which is the subject being exposed to the racial cue. And, as a result, we have (3) well defined potential outcomes and (4) a precise (and post-birth) moment of the treatment is assigned. Accordingly, extracting race-based causal inferences *is* consistent with the potential outcomes framework, provided thought has been given to the specific experimental design.

1.4.2 Observational studies with exposure to race

We can import this research design to a wide variety of *observational* contexts involving how third parties react to once they are exposed to racial signals and cues. In this sense, we could use observational data to understand how mortgage lenders react to Asian American versus white borrowers (Sen and Wasow, 2011), how juries react to Hispanic versus white death penalty defendants (Greiner and Rubin, 2010), how voters respond to political ads featuring black versus white actors, how universities respond to minority versus non-minority applicants, and how the U.S. government reacts to proposals submitted by minority-owned business in deciding to award contracts. In all

[15] Though scholars have viewed audit and correspondence studies as related, we argue that all studies employing exposure to a racial cue should be viewed as related and part of a common literature on race and causation.

of these instances, the interest lies in understanding how exposure to race changes or informs others' opinions, behaviors, or attitudes.

In the economics literature, Greiner and Rubin refer to this type of observational research as teasing apart the effects of "perceived" race (as opposed to actual race), doing so using death penalty sentencing as their motivating example. We use different terminology and draw different analogies, but the research design we suggest here is comparable to Greiner and Rubin's (2010). Nonetheless, we move away from the "perceived race" frame for two reasons. First, we think the best way to think about the "treatment" in these kinds of studies is not as perception but, instead as a signal about race. After all, in an experimental context, the researcher can manipulate the signal to which the subject is exposed but not what the subject perceives. Second, perceived race is rarely observed: what a subject perceives occurs within the confines of a mind and is generally not available to the researcher. As such, we think it more useful to focus on exposures to race and signals about race rather than perceptions of race. We opt for using this language and terminology in the observational context as well.

The caveat to this kind of study, of course, is that reliance on observational data brings with it significant drawbacks – namely that researchers lack the ability to manipulate the racial cues and signals received by the subject. In this regard, those working with observational data must worry about about satisfying the ignorability assumption, discussed above. One strategy would be to include those background variables in the analysis such that the only functional difference between the treated and control groups is that one group is exposed to minority, or other racial cues and that the other is exposed to non-minority racial cues (e.g., Sen and Wasow (2011)).

Particularly important for observational data is that the exposure framework greatly lessens problems with post-treatment bias (Greiner and Rubin, 2010). To illustrate,

suppose we are interested in whether a university accepts minority versus non-minority applicants at different rates – perhaps due to affirmative action but, perhaps also, due to discriminatory motivations. The ideal experiment here would be to mimic an audit study and create identical applicants whose profiles differ only with regard to their race. The "treatment" would be administered to the admissions officer at the time he or she reviews the application packet. Anything that happens before is solidly pre-treatment and must be conditioned on (e.g., included as variables in a regression or matching analysis); anything that happens after would be post-treatment and should not be conditioned on (Greiner and Rubin, 2010).

We conclude this part of the discussion with some practical advice for applied researchers: when possible, conceptualizing an experiment or observational study as an exposure study greatly reduces both the theoretical and practical problems associated with making race-based causal inferences. Thus, we recommend that applied researchers think carefully about whether an exposure study could provide a well-suited analogy for their research questions and hypotheses. We also note that this is a research design that is particularly apropos to questions involving racial discrimination and racial priming, thus making exposure studies an advantageous design for those in the legal, public policy, and public opinion fields.

1.5 Treatment by Manipulating an Element of Race

As we have discussed, exposure studies offer a useful framework when individuals or institutions have been presented with some signal about race. Many research questions do not, however, involve an obvious treatment by exposure to a racial cue. For example, thousands of articles have been written about racial disparities in education, health, and income. In these studies, there is generally no treatment by exposure and

no "decisionmaker." For scholars working on these and related topics, the primary research interest – and the appropriate units of analysis – lies in a particular racial or ethnic population itself.

We note that this is the category of study that most researchers think of when considering race and causality: one racial group is assigned to the treatment category (for example, African Americans), another is assigned to the control category (for example, whites), and the object of study lies in disentangling what makes the two populations different in terms of an outcome. As we noted above, however, these studies are particularly problematic in terms of having ill-defined potential outcomes and also having post-treatment bias problems.[16] In this section, we develop a framework for research designs that exploits variation *within* a racial group to extract causal inferences. In other words, this kind of research design disaggregates the "bundle of sticks" discussed above and singles out a specific element of race that can be manipulated in an experiment (or observed to vary) within a population. By identifying a mutable element of race, it is possible to identify well-defined potential outcomes and to assuage potential post-treatment bias problems.

1.5.1 Experimental studies that manipulate an element of race

As before, we begin by pointing out how this research design works with regard to experimental studies. Race, as we have noted above, has multiple components – some biological, some social, some mutable, and some assigned after birth – that form a kind of "bundle of sticks." Key to this elements of race approach is that researchers may be able to leverage the multifaceted nature of race to gain traction on different causal

[16] More practically, studies comparing two racial groups offer no obvious way to turn findings of cross-race disparities into a meaningful, implementable treatment (i.e., there is no policy that could transform members of a minority population into members of the majority).

questions.

Going back to our employment example, suppose a researcher was interested in understanding labor market outcomes for young, working-class black men – not from the vantage point of an employer,[17] but, rather, from the vantage point of the employee. An initial analysis, as we noted above, would be to take the employment outcomes and regress them on a variety of independent variables – including race, gender, age, education, job training, etc.[18] A useful step at this point would be to think carefully about which components of race the race variable is actually capturing. The researcher is likely interested in those attributes within the minority population that set it apart from the majority population and lead to different outcomes.[19]

Rather than conceive of young black men as a treated group and young white men as the control, the researcher might identify some trait that is highly collinear with being young, working-class and black and that could be manipulated experimentally, such as dialect. The researcher might then design a study in which a sample of young, working-class black men would be randomly assigned a treatment involving intensive speech coaching and training to more easily "style-shift" between black vernacular English outside of work and American standard English in all hiring and employment contexts. Comparing employment outcomes for the treated and control populations could then isolate the degree to which dialect affects labor market outcomes for young black men.[20] In essence, refocusing the study on an alternate treatment within a single

[17] For such a research question, an exposure study framework might be appropriate.

[18] As we have noted throughout, however, race is intimately linked with all of these variables (with the possible exception of gender) and is causally prior. Therefore, concerns with post-treatment bias would lead to a simplified regression with only one or two right-hand side variables.

[19] A slightly different way to phrase the question is that the researcher may want to think more precisely about the aspects of being black that might lead to employment at different rates than whites – is it racism? Differences in attitudes? Education? Different hypotheses will trigger different approaches.

[20] Such a study could also include young, working-class white men but the cross-race comparisons would only be useful for descriptive purposes, not for meaningful causal inference.

racial group could help shed light on one of the causal mechanisms that distinguishes one racial group from another.

An important point is that limiting the unit of analysis to a single racial group and reconceptualizing the treatment as being something that varies closely (but perhaps not exclusively) with race at once solves the Holland/Rubin critiques. First, re-orienting the research question to focus on another treatment – another item in the "bundle of sticks" closely tied up with social and biological constructions of race – may allow for experimental manipulation, thus avoiding the critique that there are no well-defined potential outcomes. Second, because the alternate treatment may be "assigned" post-birth, it also allows for the inclusion of all pre-treatment variables (i.e., confounders), including items like mother's education, health, nutrition, and early educational opportunities. To some extent, we are advocating treating simple biological race (as this is what the race variable now becomes) as a confounding variable that must be controlled for.[21] For applied researchers, perhaps the best strategy is to conduct a variety of within-group analyses, which have the effect of conditioning on race to solve the problem of having a lack of common support between minority and non-minority populations.

A small number of experimental studies have begun using this kind of elements of race approach. For example, one element of race – one of the sticks in the bundle – is how it affects self-assessment and ways of thinking (Steele, 1997). These psychological dimensions of race are, however, potentially mutable and amenable to experimental manipulation. Accordingly, Walton and Cohen (2011), randomly assigned a one-hour

[21] Here, the approach we are suggesting may in some instances be similar to an effects modification approach. Effects modification would be appropriate in instances where the treatment effect varies according to some different strata or subgroup (i.e., there are heterogenous treatment effects that vary systematically by subgroup). Because the impact of the alternate, non-race treatment may vary according to subgroup, comparing the results between groups may also be useful.

exercise to black and white college freshmen. The treated students watched videos and participated in other exercises that suggested all college students struggle to fit in initially but can ultimately succeed socially and academically. Compared to the black students in the control group, black students in the treated group exhibited rapid and sustained improvements in grades. The gap in academic performance between black students in the treated group and white students overall shrank by 52 percent. [22] Further highlighting the degree to which concerns about social-belonging are racialized, white students in the treatment group exhibited no significant difference from white students in the control group.

1.5.2 Observational studies that "manipulate" an element of race

Once again, what can be done experimentally can, with additional assumptions (and complications) be imported into the observational context. On this point, some observational studies have successfully leveraged additional components of race in order to extract surprising inferences. Cutler, Fryer and Glaeser (2005), for example, investigate why African Americans suffer from significantly higher rates of hypertension compared to whites. While Cutler, Fryer and Glaeser do compare blacks and whites, more telling is their comparison within black subgroups: blacks whose enslaved ancestors survived the "Middle Passage" across the Atlantic exhibited higher rates of salt sensitivity compared with blacks whose ancestors were not enslaved (i.e., recent African immigrants to the United States or the United Kingdom). The authors' working hypothesis is that salt retention – a precursor to hypertension – enabled African slaves to survive the deadly three-month sea voyage that constituted the Middle Passage. Thus, the appropriate treatment in this study was not race per se; it

[22] Walton and Cohen also report that, several years later, the black students who were treated exhibited signs of being happier and healthier as compared with the black controls.

was treatment by the Middle Passage, a finding only made clear with within-group comparisons.

Another example is provided by Nisbett and Cohen (1996), which explores why white American men in the South exhibit higher rates of violence than white men in the North. Nisbett and Cohen identify and experimentally test cultural differences they hypothesize are borne of varying immigration patterns. Where a more conventional research design might have compared rates of violence between white and black men, Nisbett and Cohen attempt to disentangle the effects of race and norms by exploiting cultural variation between Northern and Southern white men. The standard cross-race approach takes the appropriate units of analysis to be the person or person(s) of color and his or her white (or non-minority) counterpart.[23] Though cross-race comparisons are widely used in fields such as health and education, due to post-treatment bias and immutability, such comparisons are problematic when attempting to provide anything more than a descriptive analysis. In contrast to exposure studies that attempt to measure a contemporary effect of race as a signal, studies exploiting within-group variation are often attempting to identify some trait or quality assigned to members of a population at an earlier historical period. Figure 1.3 presents how clarifying the object of the study determines the unit of analysis and can help resolve whether the exposure to race or the element of race approach is appropriate.

[23] Of course, this is not the approach taken by all applied researchers. Some researchers, particularly in race and ethnic politics or in urban politics look at different measures – for example, the percent of a census tract that is minority. However, looking at minority versus non-minority populations does seem to be the general default approach.

1.6 Empirical Example

1.6.1 Who Fights in African Militias?

To illustrate these ideas, we replicated[24] Humphreys and Weinstein (2008), a recent article in the *American Journal of Political Science* that seeks to explain quantitatively the determinants of individual participation in civil war militias in Sierra Leone by using a novel closed-ended questionnaire. Specifically, Humphreys and Weinstein (2008) posit that several personal characteristics will cause an individual more (or less) likely to join one of two militia groups: the opposition Revolutionary United Front (RUF) and the government-backed Civil Defense Forces (CDF). [25] We focus our attention to the second of the seven characteristics – whether individuals are marginalized from political processes – and leave aside the other hypotheses for the moment. We also leave aside issues relating to the sampling methodology of ex-combatants and non-combatants and use the same weighting scheme used in the article.

Humphreys and Weinstein (2008) posit that political marginalization in Sierra Leone depends on two personal attributes. The first is whether an individual is a member of the Mende ethnic group,[26] and the second is whether the individual has professed support for the major Mende-backed party, the Sierra Leon's People Party

[24] This example is provisional and we welcome suggestions for additional papers to replicate. All of the data used, as well as the necessary R code, will eventually be posted to a replication archive on the Dataverse Network Project (`http://thedata.org/home`).

[25] The authors hypothesize that seven attributes may influence joining a militia group. These include (1) whether individuals are economically deprived, (2) whether individuals are marginalized from political decision making, (3) whether individuals are alienated from mainstream political processes, (3) whether individuals receive selective incentives from the militia group, (5) whether individuals would feel safer inside a fighting faction as opposed to outside of it, (6) whether other members of their community are active in the movement, and (7) whether the individual's community group has strong social structures. Taken together, these attributes explore grievances, incentives, and community and social networks.

[26] Along with the Temne group, the Mende ethnic group comprises is one of the larger and more politically successful ethnic groups in Sierra Leone and comprises roughly 30% of the population. Although the Temne-backed RUF abducted individuals from Mende villages, the recurring theme in this literature is that the RUF is backed and supported by the Temne and the CDF is backed and supported by the Mende.

(SLPP). Thus, Humphreys and Weinstein (2008) include both variables in a logit regression model that has membership in a either of the two milita group (a yes-or-no variable) as the outcome variable and a host of additional variables as controls. (That is, they take membership in the RUF as a separate outcome variable, and membership in the CDF as another outcome variable.) These controls include whether the individual is a farmer or a student, whether he or she lives in the capital, his or her gender, his or her age, and an individual measure of infant mortality. The model also includes variables designed to test Humphreys and Weinstein's (2008) other substantive hypotheses, variables that we also include in our replication. Table 1.1 presents our replication of the Humphreys and Weinstein results. Our replication is identical to Humphrey and Weinstein's to tenth decimal place (including the same number of observations), thus assuring us that we are working with the exact same data and the exact same models.

We move forward from this basic starting point by focusing explicitly on the political inclusion hypothesis, as realized by the `Mende` variable (which takes on a value 1 if Mende, 0 otherwise). Thus, we take an individual's Mende status as the ultimate question of causal interest – a somewhat distinct approach than that taken by Humphreys and Weinstein, who take a causal interest in a wide variety of variables. Our first step – essential to beginning any similar causal inquiry – is to remove from the model all of the potential post-treatment variables. Although Mende is recognized as an ethnicity as opposed to a racial group, the fact remains that being Mende is considered in Sierra Leone an immutable characteristic, and it is therefore a "treatment" administered at birth. Accordingly – and this is also mentioned by Humphreys and Weinstein (2008) – an individual's Mende status is causally prior to a host of other variables included in the model, and the inclusion of these other variables

is introducing post-treatment bias into this particular estimate. In other words, being Mende deeply affects whether a study participant lives in Freetown or not, whether he or she lives in a mud hut or not, and, most crucially, whether he or she support the Mende-backed SLPP party. All of these are, however, variables that Humphreys and Weinstein (2008) include in their model. Removing them from the model helps us estimate the effect of being Mende with less bias.[27] The results from this analysis are presented in Table 1.2.

For the RUF general membership, the significance of the Mende variable does not change – belonging or self-identifying as Mende is related positively with membership in the RUF, although the size of the effect is reduced (a move that makse sense given the RUF's status as a Temne-backed organization). The results are, however, strikingly different for the CDF. The original model (Table 1.1) shows that no causal effect can be reliably discerned on CDF membership with Mende self-identification. (The coefficient, although positive, is not statistically significant.) Once we remove the variables that could possibly introduce bias into the estimate, we see that being Mende *is* positively linked with belonging to the CDF, and that the effect is quite significant. Interestingly, the effect associated with Mende self-identification is about as strong for membership in the RUF *and* the CDF, a finding makes more substantive sense. The CDF, after all, is comprised primarily of Kamajors, a group of Mende traditional hunters, and is thought to represent Mende interests.

To explore the difference between Mendes and non-Mendes further, we focus specifically on membership in the government- and Mende-backed CDF. (This has a

[27] We may think that people from minority ethnic groups might have longer or shorter lifespans due to differences in diet and exercise habits. But we may also think that the year a person is born is also an immutable characteristic and therefore not affected by a person's race. Since opinions could differ on this, we proceed with including and not including age, and then compare and contrast the results.

simplifying effect, as we do not need to worry about forcible abductions, a common occurrence within the opposition RUF.) Table 1.3 shows the results of separate logit regressions on Mende and non-Mende populations, with membership in the CDF as the outcome variable. While we don't advocate this general strategy in all instances, it does help to isolate the importance of different traits in the different groups – and to show the effect modification associated with the ethnicity variable. Indeed, what this regression shows is that different variables are important for the different groups. For the Mende, living in mud housing (a proxy for poverty) is not a predictor of CDF membership, whereas for the non-Mende, it is. Likewise, having a friend in the CDF is not predictive, but it is for non-Mende. For the Mende group, it is being a boy or a man (as opposed to girl or woman) that is predictive of membership in the CDF.

Specifically, both Mende and non-Mende individuals are receptive to the sorts of things associated with economic grievances – poverty (by way of mud housing) and less education are more likely to lead to membership in the CDF. On the other hand, the Mende people are less susceptible to selective incentives – e.g., friendship, and, to a lesser extent, money and safety. The narrative that is supported by these results is that many Mendes already have a natural affinity for the Mende-backed CDF – not necessary through friendships, but through shared ethnicity. By contrast, for non-Mendes, for whom no pre-existing affinity exists for the CDF, monetary and social incentives are significantly more salient. This inclination is borne out by a simple interacted model, represented in Table 1.4, although we note that the interaction term is not significant.[28] Thus, while we have a hunch that there is something different about the Mende population, the results so far – using the data available – are somewhat

inconclusive.

The replication shown here has illuminated a few basic points. First, dropping post-treatment variables helped us arrive at a more reasonable estimate for Mende individuals' participation in the CDF – a result that stands in contrast to those of Humphreys and Weinstein, who do not discuss the potential role of ethnicity in predicting membership in the CDF. Basic within-group and interactive analyses helped to highlight patterns that are otherwise obscured – namely, that different mechanisms appear to be salient for Mende and non-Mende membership in the government-backed militia.

1.7 Conclusion

This article has highlighted both the pitfalls and the possibilities associated with trying to extract meaningful causal inferences about race in a quantitative framework. Most quantitative social scientists try to gain leverage on the causal impact of race by including simple dummy variables, along a standard battery of control covariates. As we note in this article, however, race presents unique challenges for quantitative scholars. First, race is resistant to manipulation and, hence, potential outcomes are ill-defined. Second, because race is "assigned" at birth, the host of characteristics that most social scientists control for (e.g., education, income, etc.) occur after the treatment is assigned and therefore potentially introduce bias into the estimate of interest. Third, an equally meaningful problem is that race is too complex to be synthesized into one neat variable.To the contrary, how a person is categorized by society or self-identifies is inextricably intertwined with tangible measures such as education, income, health, diet, economic status as well as intangible factors as culture,

[28] Mediation analyses using the `mediation` package were less conclusive.

traditions, and political and social attitudes. Thus, the introduction of a race "dummy" variable – along with attendant background covariates – oftentimes does a disservice to queries that look to make causal inferences about race-based characteristics.

The techniques described in this article may help researchers extract those kinds of inferences that capture a causal effect. First, we suggest that researchers begin by thinking whether their research design may be appropriately captured by an exposure study. This kind of research design may be particularly appropriate for those studying law and public policy, where the questions of interest frequently involve how people view and interact with racial signals and cues. Because the exposure research design avoids the pitfalls outlined above, it is serves as an extraordinarily useful, yet underused, research design.[29]

Second, when it comes to research designs focusing on minority populations themselves, we suggest that researchers think carefully about post-treatment bias issues. This is not a new warning (e.g., King, Keohane and Verba (1994)), but it carries particular urgency when issues relating to race and causality arise. Race, which is assigned in part at birth, has immutable components, which means that the host of variables that social scientists routinely control for may be determined post-treatment and could therefore introduce bias into the causal estimate. To rectify this issue, researchers interested in the causal impact of race should think carefully and what is and what is not post-treatment. Researchers are also well-advised to examine whether and to what extent including and dropping post-treatment variables influences the

[29] We note that the "exposure" and "element of race" approaches are not mutually exclusive and can be used simultaneously. As previously noted, Adida, Laitin and Valfort (2010) exploits religious variation among African immigrants to France to conduct a study in which both the correspondence technique and the manipulating an element of race technique are employed. Similarly, we could imagine a version of Bertrand and Mullainathan (2004) in which all resumes include cues that an applicant is black (i.e., was participated in the black student union in college) but that also signal variation on some dimension of race (i.e., native vs. immigrant or educated at a public vs private institution).

analysis.

Third, researchers may actually be able to focus on some alternate manipulable treatment regime that varies closely (perhaps exclusively) with race. Here, we find the analogy to the "bundle of sticks" a useful one; and even though biological race itself may not be subject to manipulation, things like name, culture, neighborhood, dialect, and diet – i.e., those variables that define the contours of racial identification – may be experimentally manipulated and observationally assessed. We do not attempt to say that such an alternate treatment may be found in all instances; rather, the takeaway is that (a) such an alternate treatment may vary closely with race, (b) may not already be included in the analysis, and (c) may explain away much of the effect previously attributed to race. Focusing on treatments other than the biological race of a subject not only solves problems with ill-defined potential outcomes, but it also forces researchers consider exactly what is being captured by the racial identification variable. Both of these are welcome considerations – both in terms of increasing statistical rigor and also in terms of increasing substantive engagement with developments in the racial and ethnic politics literature.

We also note that many possible alternate treatment regimes vary almost exclusively by race and, therefore, comparisons between whites and blacks, Hispanics and Asians, etc., may be of limited use due to problems with collinearity and a substantial (and persistent) lack of common support among key covariates. As a result, a useful way to explore whether alternate treatment regimes could be capturing some of the effects of "race" is to conduct within-group comparisons. The precise research design would, however, be idiosyncratic and will vary according to the research question and the available data. Other techniques that may help tease out potentially alternate treatment regimes include techniques developed in the effects modification literature.

We conclude by noting once more that our ultimate recommendation is that researchers interested in isolating the specific effects of race should think carefully about the appropriate experimental analogy. Doing so – while also keeping in mind what precisely he or she intends to measure by including a "race" variable – will help scholars to reconcile race and causation.

.

	RUF	CDF
Intercept	−12.48*	−26.74*
	(3.17)	(3.58)
Mud Walls	0.92*	1.61*
	(0.41)	(0.56)
Lack of Access to Education	1.09*	0.80*
	(0.30)	(0.30)
Supports the SLPP	−0.49	−0.58
	(0.67)	(0.58)
Mende	2.16*	0.58
	(0.87)	(0.65)
Does Not Support Any Party	1.29*	1.62*
	(0.57)	(0.50)
Offered Money to Join RUF	1.77*	
	(0.58)	
Felt Safer Inside RUF	−0.55	
	(0.37)	
Friend of RUF Members	0.24	
	(0.89)	
Villages Accessible by Foot or Boat Only	−0.01	0.03*
	(0.02)	(0.01)
Farmer	0.32	1.39*
	(0.56)	(0.46)
Student	0.83	1.26*
	(0.54)	(0.56)
Male	2.44*	4.06*
	(0.64)	(0.89)
Age	1.03	3.52*
	(1.20)	(1.23)
Age Squared	−0.20	−0.46*
	(0.16)	(0.15)
Freetown	−0.15	0.55
	(0.72)	(0.83)
Infant Mortality	13.52	16.85*
	(6.73)	(6.08)
Offered Money to Join CDF		3.19*
		(0.67)
Felt Safer Inside CDF		2.34*
		(0.30)
Friend of CDF Members		0.60
		(0.50)
AIC	39.82	42.66
BIC	181.10	189.46
log L	48.09	46.67

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 1.1: Logit Regression Replication of Humphreys and Weinstein (2008). Dependent variable is membership in the RUF or CDF militia groups.

	RUF	RUF	CDF	CDF
(Intercept)	-6.27^{*}	-3.49	-8.31^{*}	-7.22^{*}
	(0.75)	(1.87)	(0.84)	(1.65)
Mende	1.47^{*}	2.08^{*}	1.31	1.63^{*}
	(0.70)	(0.73)	(0.69)	(0.71)
gender	0.87^{*}	1.53^{*}	3.63^{*}	3.92^{*}
	(0.30)	(0.41)	(0.45)	(0.57)
age2		-1.46		-0.56
		(0.87)		(0.85)
age2sq		0.09		0.03
		(0.10)		(0.09)
AIC	12.83	16.66	18.72	22.61
BIC	37.76	58.21	44.63	65.79
$\log L$	5.59	11.67	2.64	8.69

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 1.2: Logit Regression Replication of Humphreys and Weinstein (2008). Dependent variable is membership in the RUF or CDF militia groups. Post-treatment variables have been removed from the model.

	Mende Only	Non Mende Only
Intercept	-21.71^{*}	-25.25^{*}
	(3.73)	(9.12)
Mud Walls	2.06^{*}	2.18^{*}
	(0.61)	(0.83)
Lack of Access to Education	0.81	0.86^{*}
	(0.52)	(0.29)
Supports the SLPP	0.00	−1.78
	(0.76)	(1.05)
Does Not Support Any Party	1.75^{*}	2.03^{*}
	(0.84)	(0.58)
Offered Money to Join CDF	3.33^{*}	3.66^{*}
	(1.09)	(1.17)
Felt Safer Inside CDF	2.20^{*}	2.72^{*}
	(0.43)	(0.37)
Friend of CDF Members	−0.82	2.09^{*}
	(0.74)	(0.65)
Villages Accessible by Foot or Boat Only	0.03	0.05
	(0.03)	(0.03)
Farmer	2.10^{*}	1.03^{*}
	(0.76)	(0.50)
Student	1.03	0.38
	(0.86)	(0.69)
Male	6.15^{*}	2.01^{*}
	(1.03)	(0.61)
Age	0.41	1.46
	(1.28)	(1.26)
Age Squared	−0.16	−0.15
	(0.15)	(0.13)
Freetown	2.55	1.47
	(1.35)	(1.48)
Infant Mortality	19.22^{*}	26.47
	(6.22)	(41.49)
AIC	38.93	35.11
BIC	148.60	149.29
$\log L$	44.53	46.45

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 1.3: Comparing models fitted on the Mende population versus on the non-Mende population. Coefficients are logit estimates (standard errors in parentheses). Outcome variable is whether an individual joined the CDF or not.

(Intercept)	−8.523*	−8.140*
	(0.850)	(1.601)
Mende	1.411	1.722*
	(0.740)	(0.780)
Male	3.474*	3.722*
	(0.435)	(0.527)
Friend of CDF Members	2.122*	2.383*
	(0.652)	(0.652)
Mende:Friend of CDF Members	−0.708	−1.133
	(0.992)	(1.025)
Age		−0.134
		(0.799)
Age Squared		−0.024
		(0.089)
AIC	22.771	26.713
BIC	65.949	87.162
$\log L$	8.614	14.644

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 1.4: Simple interacted model. Logit coefficients. Outcome variable is membership in the CDF.

2

Example I: The Effect of Race in Judicial Confirmation

Despite attempts by Presidents and by advocacy groups, federal courts in the United States are still not reflective of the U.S. population. Of the 874 federal judges in service as of 2008, only 24% were women, 10% were African American, and 7% were Hispanic (Just the Beginning Foundation, 2012). Fewer than 1% were Asian American and, even today, there are no federal judges who self-identify as Native American –

surprising given the courts' involvement in interpreting federal Indian laws. Among legal actors, politicians, and scholars, there is little dispute that the overall population of female and minority judges falls short of being representative of the general population.

Compelling explanations of why descriptive representation in the courts has been so difficult to achieve have eluded social scientists, but a possible contributor is the vetting of presumptive nominees by legal trade organizations such as the American Bar Association (ABA), the nation's largest and most prestigious lawyers' association. For example, according to recent accounts, the ABA preliminarily rejected as "not qualified" 14 of Obama's presumptive judicial nominees. Of these 14 "not qualified" candidates, nine were women and eight were racial or ethnic minorities: all had their candidacies eventually fail (Savage, 2011). The end result, as some commentators have pointed out, is that the ABA now occupies a quasi-governmental role by systematically "vetoing" certain kinds of candidates. Among liberals and racial and ethnic advocacy groups, the belief is that groups like ABA are biased against minorities and women. Among conservatives, the widespread belief is that the ABA is biased in a liberal direction, a notion that has been confirmed by a handful of empirical papers (Smelcer, Steigerwalt and Vining Jr, 2011; Lott, 2001; Lindgren, 2001). So strong is this belief that the administration of George W. Bush refused the ABA the long-standing courtesy of "pre-clearing" presumptive candidates before their names were made public and their nominations official.

This essay steps squarely into this debate. Looking at 1,652 judges confirmed since 1960 to the U.S. district courts, I find that black and female judicial nominees are significantly more likely to be awarded lower qualification ratings by the ABA, which increases the likelihood that their nominations will fail. I find that this difference persists after taking into account possible differences in educational and professional

backgrounds, age, political ideologies, and years in service. Second, and perhaps more importantly, I also find that the Bar Association scores do little to inform how a nominee will perform once confirmed onto the bench. That is, a judges who are rated as "not qualified" by the ABA are no more likely to have their opinions be overturned once they are invested than are their higher-ranked peers.

Taken together, these findings raise doubts about the process of judicial vetting, and whether an emphasis on prestige credentials (e.g., law school rank) is more important than a close look at political beliefs and partisan affiliations. This finding also calls into question the strong deference paid by political actors to the ABA's vetting process, and whether individuals who receive a "Not Qualified" rating should have their candidacies summarily withdrawn, as is currently the case (Savage, 2011). That record numbers of minority and women nominees are currently having judicial candidacies derailed by this vetting process makes this a particularly pressing issue.

This chapter proceeds as follows. Section 2 explains how the ABA assesses nominees' qualifications, focusing specifically on the selection problem that occurs when Presidents decline to move forward with poorly-rated plausible nominees. Section 3 provides an overview of the data used, which are professional and background characteristics of some 1,652 U.S. District Court judges nominated since 1960. I present the key results in Sections 5, 6, and 7, paying particular attention to sensitivity to (1) omitted variables and (2) selection bias. I conclude by discussing the implications of these results.

2.1 Evaluating Judicial Quality and Possible Bias

Once a judicial vacancy arises, the White House – working closely with the Justice Department and with the senior U.S. Senator from the state with the judicial vacancy –

develops a list of presumptive nominees via word of mouth, city bar associations, professional and political organizations, state courts, and area law firms. The short list is then forwarded to the American Bar Association's Standing Committee on the Federal Judiciary for more vetting. No rule exists mandating that Presidents must present preliminary lists to the ABA for this "pre-clearance"; nonetheless, it has been a long-standing practice followed, with key exceptions, since the Eisenhower administration. Importantly, the list of presumptive nominees is at this point confidential, and the Standing Committee members are prohibited by internal Bar rules from making the names public.

The Standing Committee[1] then begins independently reviewing each presumptive candidate's record using three criteria: (1) **integrity**, which includes "the prospective nominee's character and general reputation in the legal community, as well as the prospective nominee's industry and diligence," (2) **professional competence**, which "encompasses such qualities as intellectual capacity, judgment, writing and analytical abilities, knowledge of the law, and breadth of professional experience," and (3) **judicial temperament**, which includes "the prospective nominee's compassion, decisiveness, open-mindedness, courtesy, patience, freedom from bias, and commitment to equal justice under the law" (American Bar Association, 2009). The process by which the Standing Committee determines "integrity," "competence," and "temperament" is kept strictly confidential, and the Committee does not make any ratings public until the President confirms that the presumptive candidate will be put

[1] The Committee is composed of 15 individuals from the various federal jurisdictions (known as "circuits"). This includes the Chair of the Committee, two members from the large California-based Ninth Circuit, one member from each of the other 12 circuits. The members appointed by the ABA President for staggered three-year terms and cannot serve more than two terms (American Bar Association, 2009). Although membership is open to all ABA members, the composition of the Standing Committee has historically been white and male, with its first African American and female members appointed in 1976 and 1977, respectively.

forward as an official nominee to the Senate Judiciary Committee (American Bar Association, 2009). Thus, many Presidents have declined to pursue some number of plausible candidacies, possibly based in part on unfavorable (yet not publicly disclosed) preliminary ABA ratings. This practice could introduce substantial bias into the analysis and raises the possibility of implicit bias by the ABA when none in fact exists. I discuss this below.

The opacity of the ratings process has led to assertions that certain candidates are systematically disadvantaged. In this regard, the strongest critique has been that the ABA is biased left-ward and that ideologically conservative candidates and/or candidates nominated by Republican presidents are more likely to receive lower ABA ratings. Examining data from two administrations, for example, Lindgren (2001) finds that confirmed Bill Clinton appeals court appointees with no judicial experience had "9.7 times as high odds of getting the highest ABA rating" as similar George H.W. Bush appointees, controlling via logit regression for key differences. Although Lindgren (2001) finds no differences between nominees *with* judicial experience he does find differences in the criteria that are predictive of high ABA marks under the Clinton and Bush I regimes. (These findings were later critiqued by Saks and Vidmar (2001) on the grounds that the analysis did not include presumptive nominees, as well as District Court nominees, and could therefore be biased.) Similar results are obtained by Lott (2001), who does collect data from a handful of presumptive appeals court nominees whose names were not put forward as actual candidates. More recently, scholarly evidence in favor of a partisan bias has been put forth by Smelcer, Steigerwalt and Vining Jr (2011), which uses genetic matching to find a bias against Republican Court of Appeal nominees. They find, however, no evidence associated with either race (non-white status) or gender.

That the ABA could be partisan or biased against ideological conservatives has had substantial political ramifications. The Federalist Society, a right-leaning legal organization, publishes a newsletter entitled "ABA Watch" in which it closely monitors potentially biased treatment of conservative candidates by the Bar,[2] and numerous commentators and influential bloggers have also weighed in to provide anecdotal evidence on this issue (e.g., Whelan (2010); Mirengoff (2010); Lott (2006)). Conservative ire at the ABA crested in 2001 when George W. Bush's Attorney General, Alberto Gonzales, notified ABA President Martha Bennett that the White House would no longer allow the ABA to preview and vet the confidential short-list of presumptive candidates before the nominations became official (Gonzales, 2001). Thus, during the entirety of George W. Bush's administration, candidates were nominated regardless of the ABA's rating, and the ABA only had access to the same list of *actual* nominees that Congress, the media, and the public did. (Following the inauguration of Barack Obama in 2009, the custom of allowing the ABA to review the short list of presumptive candidates privately, before the names were made public, was resumed.) Because the Bush II era essentially circumvents the selection bias problem that plagues other administrations, I leverage these 261 nominees in my analysis below.

Comparatively less attention has been paid to the relationship between American Bar Association ratings and race and/or gender. Lott (2001) notes in passing that African American appeals court nominees – in particular African American *Republicans* – are most likely to get lower ratings, although these findings to do not go to the core of his results; Smelcer, Steigerwalt and Vining Jr (2011), on the other hand (and to their surprise), find no statistically significant relationship between race or gender and ABA qualification ratings. Anecdotally, however, the belief has increasingly

[2] http://www.fed-soc.org/publications/page/aba-watch.

been that the Bar is tilted against some of these candidates, perhaps owing to women and minorities having less "courtroom" experience and more government and/or academic experience (Savage, 2011). Obama Administration officials have, for example, have been confidentially informed that the ABA has so far "opposed 14 of the roughly 185 potential nominees the administration asked it to evaluate." Of these "nine are women – five of whom are white, two black, and two Hispanic. Of the five men, one is white, two are black, and two are Hispanic" (Savage, 2011).

This perceived negative treatment of minority candidates has, furthermore, led to tensions between the ABA and Democrats and liberal advocacy groups. (To this extent, the ABA has found itself in opposition with an unlikely combination of conservatives and liberals.) The Obama administration has declined to pursue the candidacies of some of the presumptive nominees preliminarily deemed by the ABA as being "Not Qualified," which has led to concerns about the success of its diversity initiatives (Savage, 2011). Senator Harry Reid claimed that the ABA needed to "get a new life" following its awarding of a low rating to Obama nominee Gloria Navarro (Tetreault, 2010), who was later confirmed by the Senate by a vote of 98-0. And speaking about Latina nominees specifically, Robert Raben, a member of the left-leaning American Constitution Society, wrote in a recent op-ed that

> I have not seen a single Latina nominee who wasn't either hit or slammed by some establishment group – a bar association, a leader of a not for profit, a bar leader, a judicial committee – as being 'intemperate"; lacking "seasoning"; "inexperienced", "not that bright", etc etc....There's a possibility that the entire cohort of Latina lawyers who want to be federal or state judges just don't deserve it yet, but I'm not buying it. I think there's something else going on, and I think that unearthing what may be going on within the ABA's cloistered process may help us get to the bottom of this (Raben, R., 2011).

2.2 Data

The sample of interest is 1,652 U.S. District Court judges nominated between 1960 and early 2012. (I start the clock at 1960 because the first African American district judge was confirmed in 1961, and there is no support for cross-race comparisons, and very little support for cross-gender comparisons, before 1960.) I choose the district courts as opposed to higher levels of the federal judiciary due to its size. Compared to the nine Justices serving on the Supreme Court, and to the approximately 180 judges serving on the U.S. Appeals courts (the middle level of the federal courts), approximately 700 judges serve at any given point on the U.S. District courts. This wealth of data allows us to more systematically analyze discrepancies in confirmation outcomes on the basis of sex, gender, or political affiliation. This also makes this study distinct from earlier studies – e.g., Smelcer, Steigerwalt and Vining Jr (2011); Lott (2001); Lindgren (2001) – which focus on appeals court judges.

In addition, the U.S. District Courts provide a good basis for understanding whether external qualification ratings predict judicial "performance." Of the nearly 300,000 cases per year filed in district courts, around 70,000 are appealed to the U.S. Courts of Appeals, which then reverse or uphold the lower-court judges' rulings. These rulings provide a convenient population to analyze separately: we can determine whether a lower court judge's ABA rating will be predictive of his or her reversal rate. This contrasts with the appeals courts, from which only approximately 70 cases per year are appealed to the U.S. Supreme Court.

For each of the 1,652 district court judges, I collected his or her ABA qualification rating using biographical data provided by the Federal Judicial Center.[3] The ABA currently awards three possible ratings: (1) **Well Qualified**, for which "the prospective

[3] http://www.fjc.gov.

	Not Qualified	Qualified	Well Qualified	Ex. Well Qualified	*N*
All	0.01	0.43	0.54	0.02	1652
Whites	0.01	0.41	0.56	0.03	1388
Blacks	0.01	0.57	0.41	0.00	147
Hispanics	0.02	0.56	0.41	0.01	95
Women	0.00	0.47	0.52	0.00	279
Men	0.01	0.42	0.55	0.03	1373
Democrats	0.01	0.42	0.54	0.02	726
Republicans	0.00	0.43	0.54	0.02	926

Table 2.1: Distribution of ABA Qualification Ratings for U.S. District Court Judges confirmed after 1960.

nominee must be at the top of the legal profession in his or her legal community; have outstanding legal ability, breadth of experience, and the highest reputation for integrity; and demonstrate the capacity for sound judicial temperament," (2) **Qualified**, in which the nominee "satisfies the Committee's very high standards with respect to integrity, professional competence and judicial temperament, and that the Committee believes that the prospective nominee is qualified to perform satisfactorily all of the duties and responsibilities required of a federal judge," and (3) **Not Qualified**, where the ABA has "determined that the prospective nominee does not meet the Committee's standards with respect to one or more of its evaluation criteria – integrity, professional competence or judicial temperament" (American Bar Association, 2009). Two other categories have been discontinued: (4) **Exceptionally Well Qualified**, discontinued in 1989, and (5) **Not Qualified by Reason of Age**, discontinued in 1980. Only three confirmed judges ever received the "Not Qualified by Reason of Age" rating, which was automatically awarded to individuals over the age of 60 at the time of nomination. Because so few nominees received this rating, and because this rating was deterministic, I drop it from the analysis.

A demographic breakdown of scores by race, gender, and party affiliation (by party of the appointing President) is provided by Table 3.6. Very few judges – only about 3% – were ever awarded the two most extreme categories, "Exceptionally Well Qualified" and "Not Qualified." About 43% have been awarded the second lowest category, "Qualified," with the majority of judges, 54%, being awarded the second highest category "Well Qualified." (The same is, however, not true for minority judges, more of whom were awarded the lower "Qualified" category: 57% of African Americans and 56% of Hispanics received this category.) Because so few nominees were ever awarded the two most extreme categories, and because the highest category ("Extremely Well Qualified") was abolished in 1989, I move forward by dichotomizing the qualification scheme into two categories: (1) those who received one of the highest two categories versus (2) those who received one of the lowest two categories. Dichotomizing the ABA scores in this way is routinely done in this literature, and never changes the inferences about the middle two categories – "Well Qualified" and "Qualified."

In addition to recording a judge's ABA rating, the data from the Federal Judicial Center include demographic characteristics such as age, place of birth (or death, if applicable), law school attended, past judicial experience, and a brief blurb describing the judge's previous professional experience. Because previous professional experience speaks directly to the ABA's criteria of "professional competence," I used automated content analysis to code these excerpts to indicate whether each judge had (1) legal clerkship experience,[4] (2) had worked in private practice, was (3) a full-time law professor or dean,[5] (4) worked as Congressional counsel or as (5) an attorney with the

[4] I define this as whether the judge clerked for an individual judge, as opposed to serving as a court clerk, clerk of the court, or court staff attorney, occupations that sometimes were sometimes designated by the Federal Judicial Center as "law clerk."

[5] Here, I exclude individuals who worked as adjunct or visiting professors, lecturers, or clinical instructors.

	Whites	Blacks	Hispanics	Women	Men	Dems	Reps
Ave Age at Investiture	50.43	48.55	47.66	47.93	50.50	50.59	49.65
Female	0.15	0.27	0.28	-	-	0.23	0.12
Democrat Appointed	0.40	0.71	0.49	0.61	0.40	-	-
Top 14 Law School	0.30	0.28	0.24	0.29	0.30	0.32	0.29
Private Law School	0.51	0.67	0.44	0.59	0.51	0.54	0.51
Law Clerk	0.22	0.14	0.12	0.35	0.18	0.23	0.20
Law Professor	0.05	0.12	0.06	0.07	0.06	0.07	0.05
Private Practice	0.94	0.76	0.84	0.82	0.94	0.91	0.92
US Attorney	0.09	0.03	0.05	0.06	0.09	0.06	0.10
Assistant US Attorney	0.19	0.29	0.21	0.29	0.18	0.19	0.21
Justice Dept Lawyer	0.05	0.07	0.04	0.06	0.05	0.05	0.05
Public Defender	0.03	0.10	0.14	0.06	0.04	0.07	0.02
US Magistrate Judge	0.08	0.10	0.15	0.20	0.07	0.09	0.08
US Bankruptcy Judge	0.01	0.04	-	0.03	0.01	0.01	0.01
State Judge	0.38	0.55	0.50	0.45	0.40	0.42	0.40
N	1388	147	96	279	1373	726	926

Table 2.2: Demographics of U.S. District Court Judges confirmed after 1960.

Department of Justice, and whether the judge was ever (6) a U.S. Attorney or (7) an Assistant United States Attorney. I also coded whether the judge had worked in a judicial capacity before – for example, as a federal magistrate, bankruptcy, or territorial judge, or as a state judge (both state lower court and state supreme court judge). The breakdown by race, gender, and party affiliation is reported in Table 3.3.

The Federal Judicial Center also includes each judge's gender and race or ethnicity. The racial/ethnic categorizations used by the Judicial Center are mutually exclusive, relying on self-identification, and include white, African American, Hispanic, Asian American, and Native American. (The Judicial Center therefore treats "Hispanic" as a distinct racial categorization.) Also reported is the law school and undergraduate institution each judge attended.

2.3 Methodology

Because minority and female nominees on average have differences in terms of their legal training, professional backgrounds, and judicial experience (Table 3.3), simple comparisons may mask substantial differences in these populations. To account for possible differences, I rely on matching (Ho et al., 2007). Matching allows the comparison of nominees who are identical across key characteristics. Thus, a female nominee who graduated from a Top 14 law school and who previously served as federal magistrate judge will be compared with a comparable male nominee who also graduated from a Top 14 school and who also worked as a federal magistrate judge.

This approach offers several advantages. First, matching is an effective pre-processing step that reduces dependence on statistical modeling assumptions (Ho et al., 2007). Second, and relatedly, matching effectively tests all possible ways that variables could interact with each other. We may, for example, think that the ABA might treat male and female judges differently, but only among individuals attending lower-ranked law schools. By pruning the data, matching resolves this problem and isolates the effect of a nominee being female or African American, regardless of the possible ways that other variables may be affecting one another. To implement the matching, I use coarsened exact matching (Iacus, King and Porro, 2011, 2009), which allows exact matching on key variables and coarsening and then matching approximately on the few variables that are continuous (discussed below). Coarsened exact matching has the advantage of allowing for this approximation to be as close as needed to remove biases. I also have the advantage of matching exactly – the best form of matching – on a large portion of the variables. Once nominees were matched, I took the difference in means in their ABA ratings.

As discussed below, however, we may be interested in estimating the differences in

ratings assigned over not just the subset of the population for which there is overlap in professional characteristics (e.g., the matched sample), but also over the full population of interest (e.g., all nominees). We may also be interested in how the ABA ratings differ across certain population subsets – including across different party affiliations or across different geographic jurisdictions; these may all have implications for the causal mechanism(s) behind the results. Lastly, as discussed above, we may be interested in looking at whether (and to what extent) ABA qualification ratings could be useful predictors of a judge's performance once confirmed on the bench. Thus, I at times fit logit models, in most instances controlling for the same variables used in the matching. In addition, because the coefficients obtained using a logit link function can be difficult to interpret, I present predicted probabilities throughout. The substantive results of these models reinforce the results from the matching.

At all times I match on, or control for, key personal characteristics of the U.S. District Court nominees, including whether the nominee (1) was a former law clerk, (2) had ever served as a United States attorney or as an Assistant United States attorney, (4) had worked in the Solicitor General's Office (as a Deputy or Assistant Solicitor General), (5) had ever served as a state judge (either as a state supreme court or state lower court judge judge), (6) had ever been a former federal judge (e.g., magistrate, territorial, or bankruptcy judge), (7) had worked as a full-time law professor or law school dean, (8) had experience as an attorney in private practice, or (8) had ever been a public defender.[6] I also match the judges' law school 2001 U.S.

[6] Making causal claims pertaining to immutable characteristics is particularly thorny because the "treatment" (e.g., race or ethnicity, gender) is assigned at birth, rendering (1) experimental analogies ill defined and (2) nearly all control variables post-treatment (Greiner and Rubin, 2010; Sen and Wasow, 2011). Here, I conceptualize the treatment not as being the nominee's race or gender, but the *exposure* of American Bar Association's Standing Committee on the Federal Judiciary to the nominee's immutable gender, race, or ethnicity (Sen and Wasow, 2011). Thus, a well defined experiment would be taking identical nominees (with identical profiles) and randomly assigning the "race" associated with the nominees – similar to what is done in audit studies in public health, housing, and labor economics. Conceptualizing the observational

News & World Report rankings, dividing them into six categories: (1) elite law schools in the "Top 14," (2) other law schools in the Top 25, (3) law schools ranked between 26-50, (4) law schools ranked between 51-76, (5) law schools ranked 76-100, and (6) law schools ranked outside of the top 100. (These are admittedly a somewhat rough measure for judges attending law school in the 1960s and 70s.) I also include a dummy variable for whether the law school was public or private and include in the analysis the nominee's age, coarsening to create four age cohorts: (1) 30-40, (2) 40-45, (3) 46-55, and (4) 55+.

In terms of political ideology, I include two key variables (where appropriate). First, I match on, or include dummy variables for, the President who nominated the judge. This also has the effect of conditioning on administration idiosyncrasies and possible fluctuations due to external historical or social trends. Second, for those nominees who went on to be confirmed and invested (i.e., those who went on to become district court judges), I use the judges' judicial common score (Boyd, 2011; Epstein et al., 2007; Giles, Hettinger and Peppers, 2001; Poole, 1998), which relies on the party of the appointing President or, if the party of the appointing President coincides with that of the senior senator of the nominee's state, the common score of the senior senator.

A summary of characteristics post-matching is given by Table 2.3. This matched sample of judges is, as expected, slightly different than the original pre-matched sample (the first column of Table 2.3 and as well as Table 3.3) but certainly not fundamentally atypical. Very few of the matched judges had experience working as magistrates or bankruptcy judges, as law professors, or as Assistant U.S. Attorneys, a testament to the

study in this way highlights that the moment of "treatment" happens when the nomination packet is assembled and initially presented to the Bar Association. Lastly, it is also important to note that conceptualizing the "treatment" as being assigned at birth does not actually affect the core findings: black, female, and Hispanic nominees still receive lower ABA scores even when no statistical controls are included (Table 3.6).

	All	Whites	Blacks	Women	Men	Dems	Reps
Female	0.17	0.00	0.00	1.00	0.00	0.03	0.03
Democrat	0.44	0.61	0.61	0.72	0.72	1.00	0.00
Top 14 Law School	0.30	0.33	0.33	0.69	0.69	0.36	0.36
Law Professor	0.06	0.00	0.00	0.03	0.03	0.01	0.01
Private Practice	0.92	1.00	1.00	0.97	0.97	0.99	0.99
Assistant US Attorney	0.20	0.11	0.11	0.19	0.19	0.06	0.06
Justice Dept Lawyer	0.05	0.07	0.00	0.06	0.07	0.04	0.03
Law Clerk	0.21	0.17	0.17	0.31	0.31	0.11	0.11
Magistrate Judge	0.09	0.00	0.00	0.03	0.03	0.01	0.01
Bankruptcy Judge	0.01	0.00	0.00	0.00	0.00	0.00	0.00
State Judge	0.41	0.67	0.67	0.34	0.34	0.40	0.40
Ave Commission Year	1988.37	1988.95	1988.56	1997.31	1997.21	1983.62	1985.89
N	1652	24	18	32	37	362	499

Table 2.3: Pre-matching (for all judges) and post-matching characteristics, for (1) blacks compared to whites, (2) women compared to men, and (3) Democrats compared to Republicans.

small number of such individuals in the population of judges at large. In addition, the matched sample has (in most instances) a greater proportion of individuals who attended a Top 14 law school, whose careers were spent in private practice, and who were nominated by Democrats. Lastly, the average commission year fluctuates somewhat from the overall sample, reflective of the fact that certain candidates (e.g., women) are nominated more frequently in later administrations.

2.4 Predictors of ABA Ratings

I begin by showing how various judicial characteristics are predictive of the ABA ratings awarded. Here and in subsequent analyses the outcome variable is whether the nominee was highly rated by the ABA, receiving either a "Exceptionally Well Qualified" or "Well Qualified" rating. Thus, Figure 2.1 shows the relationship between key professional characteristics and whether a nominee earned one of the higher ABA ratings versus one of the lower ones. (Inferences about the two middle categories do

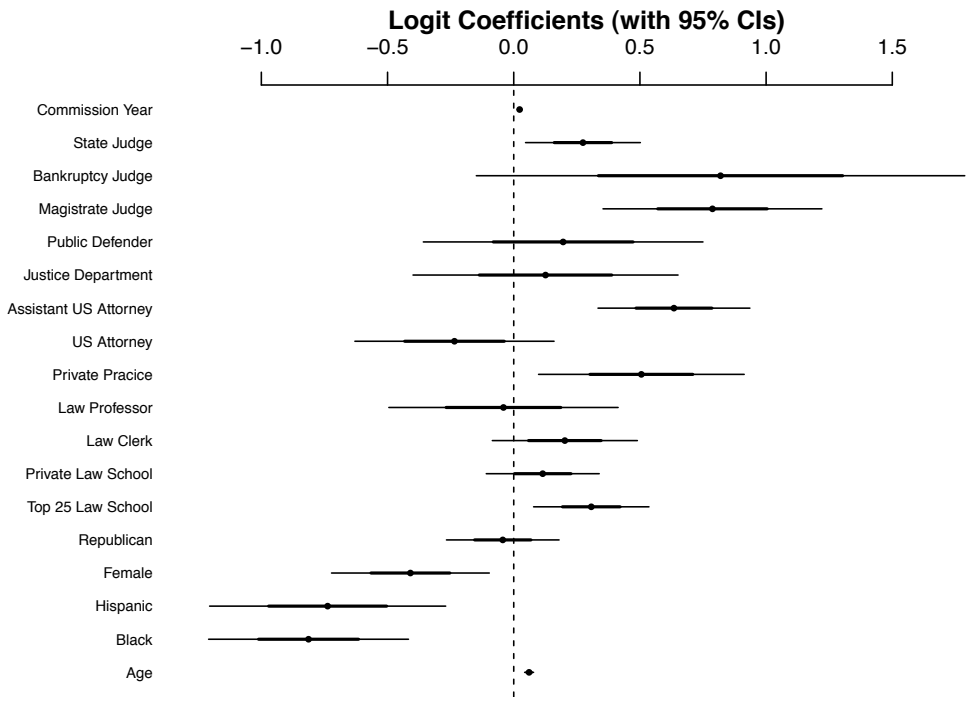

Figure 2.1: Logit regression results. Outcome variable is Exceptionally/Well Qualified rating (versus Not Qualified/Qualified rating) from the ABA. Solid dots represent point estimates; lines present one- and two-standard deviation intervals.

not change using an ordered logit specification.) I also include dummy variables for race or ethnicity (with whites comprising the baseline group), gender, and appointment by a Republican, which are the variables of interest in the analysis. Other controls include the age of the nominee and the rank cohort of the nominee's law school. (Dummy variables for the identity of the appointing President do not change the results; these are reported in Table 2.11 in the Appendix.)

As the results from Figure 2.1 suggest, certain traits are positively linked with earning a high ABA score. For example, individuals who have previous judicial experience (e.g., previous experience as a state judge, a U.S. Bankruptcy Judge, or a U.S.

Magistrate judge) are more likely to receive the higher two ratings. Other characteristics that are linked with higher scores include whether the judge attended a Top 25 Law School, spent time in private practice, or served as a U.S. Attorney. Age (here measured in years at time of commission) is also positively associated with receiving a higher ABA score. Two other characteristics – whether the judge was a former law clerk and whether the judge attended a private law school – are also positively linked with higher ABA ratings, but fall just shy of statistical significance. Thus, we can identify that prestige (e.g., high law school rank) matters, as does practical experience – including private practice experience and judicial experience.

Three traits are negatively linked with receiving low ABA ratings. These include a judge being (1) female, (2) African American, or (3) Hispanic. Effects for all three are statistically significant. (There were insufficient numbers of Asian American or American Indian judges to make meaningful inferences about these groups.) A fourth variable of interest, a judge being Republican, does not seem to have much of a relationship with the rating awarded, and it is not statistically significant under any model specification. Thus, at a preliminary level, we see some evidence that women and racial ethnic minorities receive lower ABA ratings than men and white nominees, even after controlling for key judicial characteristics, and no support for ideological differences based on the party of the appointing President.

2.5 ABA Ratings, Racial Minorities, and Women

The results presented in Figure 2.1 suggest that racial/ethnic minorities and women are receiving lower scores, even after controlling for key judicial characteristics. However, it could be the case that these results are model dependent, possibly obscuring the true role played by race/ethnicity and gender. Relatedly, it is also possible that the model is

making predictions outside of the support of the data – e.g., in instances where there are no substantially comparable whites and blacks, or men and women. Thus, I now turn to matching to more closely investigate the potential role played by race and gender.

Unfortunately, because there are so few Hispanics, and nearly no Asian Americans (and no Native Americans), I focus the race/ethnicity part of the analysis on African Americans, here compared to whites. In each instance, unless noted otherwise, I first match on the relevant personal and professional characteristics, including (1) judge gender (or race in the case of women, discussed below), (2) the identity of the appointing president, (3) age (using four age cohorts), (4) state judge, (5) U.S. Attorney, (6) Assistant U.S. Attorney, (7) Solicitor General Assistant or Deputy, (8) Federal Magistrate or Bankruptcy Judge, (9) law professor, (10) private practice experience, (11) public defender experience, (12) law clerk experience, (13) law school rank, and (14) ideology (as measured by the nominee's judicial common score). Next I calculate the difference in means in the two populations (black and white nominees) in terms of the ABA rating awarded.

Results from after matching on these key characteristics are presented in Table 2.4. Looking at African Americans, for example, an estimate of -34% indicates that black judicial nominees are on average 34% *less* likely to receive a high rating from the ABA than are professionally similar whites nominated by the same Presidents, a difference that is also statistically significant at the 5% level with 95% confidence intervals of -55% to -8%. (Conversely, the same analysis also results in African Americans being 34% *more* likely than similarly situated whites to receive the lowest two ratings, and that difference is also statistically significant.) Different coarsening and including other professional factors into the analysis never change the direction or even rough magnitude of the results.

	Prob Change in Receiving High Rating	95% CI
African Americans	-0.34	(-0.55, -0.082)
Women	-0.24	(-0.43, -0.036)

Table 2.4: Change in probability, after matching, of receiving one of the highest two ratings from the ABA.

The results attenuate slightly for female nominees. For women, I match them to men across the same characteristics as before; the one exception is that instead of matching on the nominee's gender, I match on the nominee's race or ethnicity (so as to hold that constant). The results, presented in the second row of Table 2.4 demonstrate that women are, on average, 24% less likely than similarly situated men to receive a high rating from the ABA. Women are also more likely to receive the lower two ratings awarded by the American Bar Association ("Not Qualified" and "Qualified"). Both findings are statistically significant.

Sensitivity to Omitted Variables

Although I match on, or otherwise take into account, a substantial number of factors that could possibly influence the scores awarded, it is clearly possible that (1) we do not have access to the full breath of information available to the ABA's Standing Committee on the Federal Judiciary or that (2) some of the information used by the ABA is inherently qualitative in nature and not included in the Federal Judicial Center's amalgam of data.

To gain some traction over the possibility that unobserved covariates are driving the results presented in Table 2.4, I use a method of sensitivity analysis described by Rosenbaum (2002), implemented in R using the `rbounds` package developed by Keele (2010). This sensitivity analysis works roughly by hypothetically "increasing" the level

	Post-Matching Coefficient	*p*-value	Γ Statistic
African Americans	-1.63	0.0022	1.70
Women	-1.10	0.006	1.25

Table 2.5: Original post-matching logit coefficient estimate, exact *p*-value under no confounders, and Rosenbaum sensitivity analysis Gamma value.

of unobserved covariate(s) in the "treated" population (e.g., racial and ethnic minorities, women) until the results are no longer significant. Thus, the sensitivity analysis gives us an estimate of the size of the bias (denoted as Γ) that must be present in these populations in order for the results to be called into question. For example, a result of $\Gamma = 1.2$ for African American nominees means that there must be 20% more of some unobserved trait among the African American nominees for the results to lose significance. Although there is no firm agreement in the literature about the minimum Γ value for observational studies, anything above $\Gamma = 1.5$ appears to indicate substantial insensitivity to unobserved confounders.

The results are presented in Table 2.5 and demonstrate that (by observational standards) the results are actually fairly insensitive. In order to yield the results insignificant, some trait would have to be present in the African American judge population approximately 1.70 times as often as in the white population. Given that the analysis already controls for clerkship experience, professional experience, quality of legal education, and previous judicial experience, this seems unlikely. It could be the case that, for example, African American judges are 70% less likely than white judges to have been on their school's law review or to have graduated as members of the Order of the Coif, a law school honors society. Controlling for the law school's ranking and for subsequent judicial experience (for which such metrics might be predictive), again, makes this somewhat unlikely. For women, the results are more sensitive, a result

consistent with the smaller treatment effect for this group (Table 2.4). In order to yield the results insignificant, women nominees would have to have some treatment approximately 1.25 times as often as male nominees.

Sensitivity to Selection Bias

As noted, the ABA makes public its qualification ratings only for those individuals who were eventually nominated by the White House and whose candidacies advanced to the Senate Judiciary Committee; that is, ABA qualification scores are available only for *actual* nominees, not *presumptive* nominees (American Bar Association, 2009). In addition, the Federal Judicial Center collects the ABA scores and previous professional and judicial experience of those nominees who were actually nominated by the White House, confirmed by the Senate, and invested as U.S. District Court judges. Thus, both the Federal Judicial Center data (and additional data collected by Zuk, Barrow and Gryski (2009)) systematically exclude ABA ratings of (1) individuals whose candidacies were dropped during the ABA's "pre-clearance" stage, and (2) individuals who were actually nominated by the White House but whose nomination eventually failed or was withdrawn.

Although not publicized, anecdotal evidence suggests that the actual number of failed presumptive nominees appears to be quite small, somewhere around 3-5 per four-year term.[7] A significant concern is, however, that not including these individuals in the analysis could bias the results. For example, it could be the case that Presidents starting with Jimmy Carter were eager to appoint minority judges, perhaps in order to

[7] According to Lott (2001), "three potential nominees were said to have been advised that they would get a 'not qualified' rating during Bush I and nine potential nominees fell into this category for Reagan"; Bush II did not submit names for ABA "pre-clearance," while Obama, an exception, has had about 14 nominees whose names have not moved forward due to receiving a poor ABA mark (Savage, 2011). The identities of these failed presumptive nominees is strictly confidential.

President Name	Whites	African Americans	Hispanics	Women	*N*
Barack Obama*	0.72	0.17	0.11	0.47	108
George W. Bush	**0.82**	**0.07**	**0.11**	**0.21**	**261**
William J. Clinton	0.76	0.18	0.06	0.29	305
George H.W. Bush	0.89	0.07	0.04	0.2	148
Ronald Reagan	0.93	0.02	0.05	0.08	290
Jimmy Carter	0.78	0.15	0.07	0.15	195
Gerald Ford	0.91	0.06	0.02	0.02	49
Richard M. Nixon	0.96	0.03	0.01	0.01	178
Lyndon B. Johnson	0.92	0.05	0.03	0.02	115

*Recorded by FJC as of April 3, 2012

Table 2.6: Racial/ethnic and gender distribution of judicial nominees by President (Johnson through Obama administrations)

increase more rapidly the proportion of black and women judges on the courts. Under such a scenario, it is quite possible that Presidents who had their "short lists" vetted by the ABA would move forward by officially nominating "Not Qualified" minority or female candidates to the full Senate, while declining to move forward the nominations of "Not Qualified" white or male candidates. The same is true for individuals who were actually nominated by the White House but were rejected by the Senate or withdrew their nominations. In that context, the bias would come from the Senate Judiciary Committee being more likely to reject "Not Qualified" white or male nominees while pushing forward "Not Qualified" minority or female nominees, perhaps due to concerns about diversity and/or not wishing to appear biased. The observable implications of both would be that the ratings awarded to confirmed candidates by the ABA would appear skewed against women or minority candidates, even though there would be no bias associated with the ratings process itself.

George W. Bush Nominees. As noted, George W. Bush declined to allow the ABA to evaluate presumptive nominees in advance of their nominations (Gonzales, 2001). Thus, during the years 2001 to 2008, we have *all* of ABA scores awarded, which avoids the selection bias problem present elsewhere in the data. A politically awkward situation is therefore empirically quite useful.

Because only 18 African American, 27 Hispanics, and 54 women judges were nominated during the Bush II years (Table 2.6), I use parametric methods instead of matching. Table 2.7 shows results from a logit regression including race, gender, and a variety of professional and educational characteristics[8] where the outcome variable is whether the nominee (here, the actual nominee) received either (1) high rating ("Well Qualified"), or (2) a low rating ("Qualified" or "Not Qualified"). I include one model with dummy variables for the district of origin and one model with race and gender interacted. The results are only fleetingly significant, owing to the small number of racial/ethnic minorities and women. However, the results from the Bush II years are largely consistent with the results seen before: although not significant, the model coefficients are suggestive of African American, Hispanics, and women nominees being less likely than, respectively, whites and men to receive the higher two ABA ratings. For women and for Hispanics the effects are no longer significant; for African Americans, they are significant when the effect is allowed to vary across district court jurisdiction (Model 2).

Artificially constructed presumptive candidates. The fact that George W. Bush nominated only 18 African Americans to the district courts hampers the ability to extract meaningful estimates about his terms. To provide additional context, I

[8] I do not include the judicial common scores measuring ideology because about half are missing and including them could potentially introduce bias.

	Model 1	Model 2	Model 3
(Intercept)	−4.83*	15.51	14.60
	(1.49)	(5318.04)	(5321.65)
Age	0.10*	0.07	0.08
	(0.03)	(0.04)	(0.05)
African American	−0.32	−2.18*	−2.36
	(0.63)	(0.99)	(1.24)
Hispanic	−0.03	−0.77	−0.08
	(0.53)	(0.79)	(0.95)
Female	−0.25	0.00	0.52
	(0.38)	(0.62)	(0.81)
Top 25 Law School	−0.00	−0.05	0.08
	(0.36)	(0.58)	(0.60)
Private Law School	0.61	0.37	0.43
	(0.32)	(0.54)	(0.55)
Law Clerk	−0.20	−0.82	−0.76
	(0.34)	(0.63)	(0.64)
Law Professor	0.55	−0.01	0.12
	(0.87)	(1.50)	(1.53)
Private Practice	0.25	−0.65	−0.48
	(0.48)	(0.84)	(0.91)
US Attorney	−0.21	−0.13	−0.25
	(0.57)	(1.17)	(1.18)
Assistant US Attorney	1.43*	1.92*	1.99*
	(0.43)	(0.75)	(0.78)
Justice Department	−0.86	−1.07	−0.88
	(0.68)	(1.44)	(1.50)
Public Defender	0.65	1.46	1.89
	(0.83)	(1.77)	(1.96)
Federal Magistrate	0.82	1.18	1.20
	(0.49)	(0.85)	(0.85)
State Judge	0.10	0.04	0.08
	(0.33)	(0.54)	(0.55)
African American*Female			0.51
			(1.98)
Hispanic*Female			−1.94
			(1.39)
District Dummies		✓	✓
N	257	257	257
$\log L$	−92.41	205.96	213.17

Standard errors in parentheses
* indicates significance at $p < 0.05$

Table 2.7: Logit regression results, George W. Bush nominees. Outcome variable is receiving a high ABA rating.

therefore artificially replicate the possible pool of presumptive candidates. Using the fact that we know the rough number (if not the identities) of the presumptive candidates rejected by the ABA, I include in the data generated observations designed to present the worst possible scenario for the key results.

To create the artificial set of observations, I generated several "presumptive nominees" per President. I did so by assuming that 8% of each President's nominees had their candidacies fail at the pre-clearance stage. (The exception here is George W. Bush.) This is significantly higher than the actual number, which appears to be around 2-4% (Lott, 2001), but closer to Barack Obama's very high average of 7.6% (Savage, 2011). The most bias would be introduced when Presidents fail to move forward poorly rated whites: not moving these individuals forward (while moving forward poorly rated minorities and women) would result in a skewed post-selection sample. Thus, I initially create an artificial sample of 120 "failed nominees" who are both white, young, and poorly qualified by the ABA, and I assign them those covariates least linked with higher ABA ratings (including no prior judgeships, law clerkships, or private practice experience).

After including them in with the original data, I re-ran the key analyses, which are presented in Table 2.8. The original results are insensitive to their inclusion, particularly for African Americans and for Hispanics, for whom the relationship to high ABA scores is still negative and significant under any model specification. For women, the results are still significant once we allow the effect to vary across Presidential administration (Model 2). The results are therefore not broken, even under these extreme assumptions. In addition, incrementally increasing the number of "presumptive judges" in the artificially created set (dropping covariates), shows that the fraction of presumptive nominees dropped due to the selection bias out of the total

	Model 1	Model 2	Model 3
African American	−0.62*	−0.70*	−0.74*
	(0.19)	(0.20)	(0.23)
Hispanic	−0.53*	−0.73*	−0.75*
	(0.23)	(0.24)	(0.27)
Female	−0.14	−0.31*	−0.34*
	(0.15)	(0.16)	(0.17)
African American:Female			0.15
			(0.44)
Hispanic*Female			0.11
			(0.52)
President Dummies		✓	✓
N	1749	1749	1749
$\log L$	−1005.10	−946.22	−940.15

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 2.8: Logit coefficients generated when including 120 generated observations to reflect unknown "presumptive nominees." Controls for professional experience, age, and education not shown.

number of nominees would have to be 10% to break the results for women, 11% to break the results for Hispanics, and 15% to break the results for African Americans – nearly twice the rate recently reported for the Obama administration (Savage, 2011). Thus, we have little evidence that the results seen here are driven by this particular selection bias problem.

I note that these sensitivity tests do not yet address the second source of bias identified – actual nominees whose nominations were withdrawn. These include around 200 publicly named nominees whose candidacies were (1) withdrawn by the White House or by the nominees themselves, (2) rejected by the Senate, or (3) withdrawn due to the nominee's death. For these individuals, I have begun collecting the same information as those who were eventually confirmed and invested – e.g., data on their legal education, previous professional experience, any judicial positions (if

any), and their ABA rating. The data were not ready by time of writing, but their relatively small number (200 compared to 1,600) means that the bias resulting from their exclusion is likely limited, and a preliminary check is provided by the same analysis presented in Table 2.8.

2.6 ABA Ratings and Party Bias

A remaining issue is whether, as many have alleged, there is an ideological bias to the scores assigned by the American Bar Association, one that discriminates against ideological conservatives and Republicans. Here, at least three empirical studies – Smelcer, Steigerwalt and Vining Jr (2011); Lott (2001); Lindgren (2001) – have found such an effect looking at the U.S. Courts of Appeals. Although there is good reason to think that appointments to the U.S. Courts are more likely to be driving by political concerns, a politically biased vetting process would likely also extend to the hundreds more appointments to the U.S. trial courts.

To test this, I focus on two measures of partisanship. The first is the party of the appointing president, while the second is the nominee's judicial common score. The judicial common score takes advantage of "senatorial courtesy," the longstanding practice of Presidents to consult U.S. Senators on judicial vacancies in their home states. Thus, the judicial common scores are based on either the (1) the ideological common score of the appointing president or, in instances where the President and the senior senator of the judge's home state are the same, (2) the common score of the senior senator. In addition, because the sentiment of discrimination comes largely from Republican administrations, I more closely analyze judges confirmed during those time periods.

As before, I match on numerous variables: race, gender, age (broken into cohorts),

state judge, U.S. Attorney or Assistant U.S. Attorney, previous judicial experience (as a state judge or term-limited federal judge), law school rank (and whether the law school was private or public), clerkship experience, private practice experience, law professor, and public defender. Because the purpose of this analysis is to detect differences across partisan appointments, I do not match on the identity of the President making the appointments. I also do not match on the judicial common score. (Characteristics of the post-matched population is provided by Table 2.3.) As before the outcome variable of interest is receiving a high rating ("Exceptionally Well Qualified" or "Well Qualified") from the ABA.

Post-matching logit results are presented in Table 2.9. I include at four specifications: looking at party of the appointing president (Model 1), ideology (Model 2), both (Model 3), and both interacted (Model 4, which I use to test the idea that ideologically conservative nominees might be awarded lower scores by the ABA, but mostly when they are appointed by Republican administrations). As the results in Table 2.9 show, however, there is no statistically significant relationship between any of the possible treatments and receiving a high or low ABA score. For party of the appointing President, the relationship is substantively weak and falls shy of statistical significance (despite the substantial post-matching sample size), and this result dovetails with the parametric regression results presented in Figure 2.1 as well as Table 2.11. The same is true for looking at ideology: we simply cannot rule out that there is no relationship between more conservative ideology and receiving a lower score. An interaction between the two is also not significant. Thus, I see no evidence that candidates nominated by Republican presidents or who are more conservative are systematically awarded lower scores than similarly pedigreed candidates nominated by Democrat presidents or who are more liberal. This is a finding that stands in contrast

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.43*	0.35*	0.46*	0.45*
	(0.11)	(0.07)	(0.15)	(0.22)
Republican	−0.08		−0.21	−0.20
	(0.14)		(0.26)	(0.28)
Judicial Common Score		−0.06	0.19	0.13
		(0.21)	(0.37)	(0.70)
Judicial Common Score * Republican				0.08
				(0.82)
N	861	781	781	781
AIC	1175.27	1083.71	1084.81	1086.83
BIC	1213.34	1120.99	1140.74	1161.40
$\log L$	−579.64	−533.85	−530.40	−527.41

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 2.9: Effect of party of appointing presidents, post matching. Logit coefficient estimates, with outcome being receiving a high ABA score. Larger (positive) judicial common scores indicate increased conservative ideology.

with earlier literature on appointments made to the Courts of Appeals, an aspect which I discuss below.

2.7 ABA Ratings as Predictors of Judicial Performance

The results presented so far call into question the impartiality of American Bar Association qualification ratings for African Americans and for women, while complicating earlier conclusions about their partisanship or ideological bias. I now turn to a separate question, which concerns the utility of ABA scores. Given how much effort goes into calculating these scores, and deference paid to them by political actors, we would expect that ABA ratings serve some useful function or signal. In the nominations context, the greatest utility would be if ABA scores somehow predict how judge will fare once invested onto the bench – that is, how frequently the cases they

write are reversed or upheld.

I note at this point that simple reversal is not an universally agreed-upon measure of judicial "quality" or "performance," which are inherently slippery concepts, and that a lively normative debate is ongoing about whether, and to what extent, judges should be held to performance standards. When it comes to the lower courts, however, there is some agreement that certain judges systematically produce opinions of poorer quality or poorer legal reasoning, which in turn are more consistently reversed. (This assertion might be ring less true in the higher courts, where the discretionary nature of review means that high-quality opinions may be turned over due primarily to political considerations.) Thus, if ABA scores are useful predictors of anything, it should be of a district judge's reversal rate.

Case outcomes data. To test this possibility, I use an extant database of cases by Songer, Kuersten and Haire (2007). These cases represent a randomly selected subset of 12,519 cases appealed from the U.S. District Courts to the U.S. Courts of Appeals between 1960 and 2002.[9] For each case, I have data on the ultimate decision by the appeals panels. I operationalize this as being dichotomous: the appeals panel either upholds the lower-court opinion or reverses it, either in its entirety or in part. (More sophisticated measures of higher-court outcomes do not meaningfully affect the results.) In addition, I also have the identity of the U.S. District Court judge who wrote the lower-court opinion, including his or her ABA rating. I use both pieces of information information to calculate for 1,044 district court judges his or her reversal rate over this time period. (Again, I limit the population to judges who were confirmed after 1960.) A histogram of the distribution of judges' reversal rates is presented in

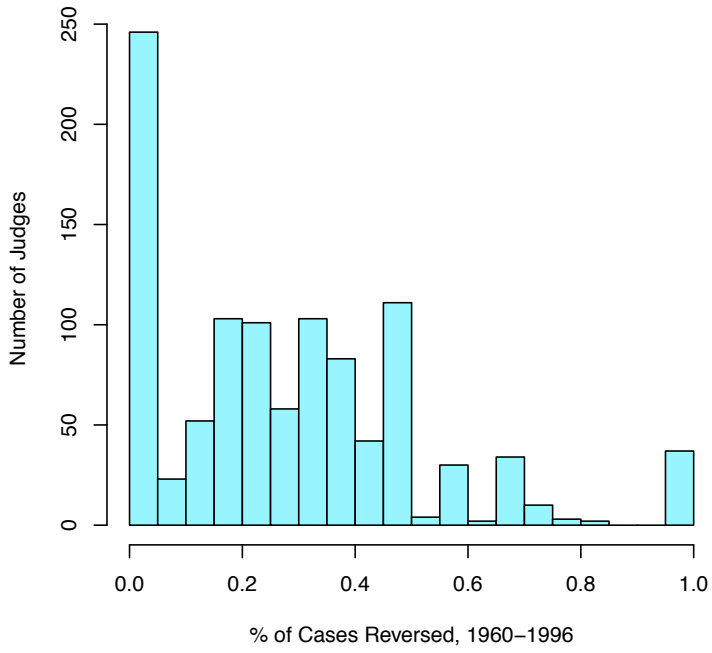

Figure 2.2: Reversal rates for U.S. District judges confirmed after 1960 (using cases decided on appeal from 1960 to 2002).

Figure 2.2.

Results. Table 2.10 presents the results from a simple OLS model where the key explanatory variable is the score awarded to the judge, with the outcome become the judges' reversal rate on cases decided between 1960 and 2002. I either dichotomize the ABA ratings categories ("highly qualified" versus "poorly qualified," Model 1) or include the full spectrum of ABA ratings, taking "Not Qualified" as the baseline (Model 2). Thus, Model 2 explicitly tests whether "not qualified" candidates are indeed substantively different than those the ABA deems "above the bar." What Table 2.10

[9] Data more recent than 2002 were unavailable at the time of writing.

	Model 1	Model 2	Model 3
(Intercept)	0.30*	0.26*	0.40*
	(0.01)	(0.09)	(0.05)
High ABA Rating (yes or no)	−0.02		−0.01
	(0.01)		(0.02)
"Qualified"		0.04	
		(0.09)	
"Well Qualified"		0.02	
		(0.09)	
"Exceptionally Well Qualified"		0.06	
		(0.10)	
Republican			−0.03*
			(0.02)
Law Clerk			0.01
			(0.02)
State Judge			0.02
			(0.02)
Top 25 Law School			−0.03*
			(0.02)
Circuit Dummies			✓
N	1044	1044	1043
R^2	0.00	0.00	0.08
adj. R^2	0.00	−0.00	0.04
Resid. sd	0.24	0.24	0.24

Standard errors in parentheses

* indicates significance at $p < 0.05$

Table 2.10: OLS regression of a judge's reversal rate (for cases decided on appeal between 1960-2002) on ABA qualification ratings.

demonstrates, however, is that ABA ratings under neither schema are predictive of a judge's reversal rate. Indeed, not only are the coefficients close to zero and statistically insignificant, but the R^2 of the regression is close to zero as well.[10]

[10] Stylizing the outcome variable not as the judge's overall reversal rate, but as the probability that *an individual case* will be reversed (in whole or in part) also results in null results with regard to the lower court judge's ABA score when judge-specific random effects are included along with dummy variables for Circuit or year.

Because ABA scores ostensibly collapse the entirety into nominee's qualifications into one easy-to-digest value, they should be predictive of judges' reversal rates without additional controls. Nonetheless, it is likely that norms about reversal vary across jurisdiction and through time, which could reveal the ABA scores's predictive value. It is also possible that reversal varies independent of ABA score according to law school rank, previous judicial experience, etc., and that the scores provide additional value added on top of these other kinds of signals. I therefore include additional predictors in the analysis (Model 3), including dummy variables representing the judicial district in which the appeal arose ("Circuits"). The results from this specification (Table 2.10, Model 3), show that there is still no relationship between ABA scores awarded to a judge and his or her rate of reversal. On the other hand, we do see a negative relationship between graduating from a Top 25 top law school and reversal rate, as well as a negative relationship between being Republican and being reversed. Dummy variables for the U.S. Appeals Courts are also significant (not shown), which suggests that norms about reversal may vary from jurisdiction to jurisdiction. In sum, there is little evidence that the score received by a judicial nominee in any way predicts how successful he or she will be in avoiding reversal.

2.8 Conclusion

The contributions of this chapter are threefold. First, the results show no differences between nominations made by Democrats and Republicans or among nominees with different ideological common scores – a finding in contrast to previous literature. Although more work needs to be done, the reason may be rooted in the different roles played by U.S. District and U.S. Appeals courts. The latter perhaps have a stronger reputation for being partisan; not only are their nominations taken more seriously by

political actors, but decision making on the appeals courts has been shown to be closely linked with party affiliations (Sunstein et al., 2006). Thus, it would not be surprising to see partisanship play a key role in the nominations of appeals court judges and less so for judges at the district level.

Second, although the results show no differences between Democrat and Republican nominees, my findings suggest that women and minority judicial candidates systematically receive lower qualification ratings from the ABA. This is the case both a priori and also when using matching or other controls to compare candidates who are similar or identical across key professional, educational, and political characteristics. The results also appear robust to potential selection issues arising from the practice of Presidents "pre-clearing" potential nominees with the ABA. The effect is present both in Republican and Democrat administrations, and sensitivity analyses suggest that it is not being driven by variables omitted from the analysis.

One way to understand these results is that the law is a prestige-oriented profession – one driven by high-status accomplishments and the general appearance of success. To this extent, it is not surprising that law school rank, previous legal clerkship experience, private practice experience, and public defender experience are predictive of the kind of ABA rating a nominee will receive. On the other hand, in instances where prestige, power, and appearances of success are paramount, we might also not be surprised that women and minorities may be systematically disadvantaged.

Third, and perhaps most importantly, the findings show that ABA ratings are not predictive of judges' ultimate performance once they are confirmed. Indeed, the analysis here demonstrates that nominees designated as "Not Qualified" to serve by the ABA have reversal rates that differ little from those awarded the stellar "Exceptionally Well Qualified" and "Well Qualified" ratings. This fact is surprising given that the ABA

ostensibly takes into account those aspects which would make for a strong judicial career – both objective criteria like law school attended, and also subjective criteria such as "temperament," "competence," and "integrity."

Ultimately, however, the results presented here call into question the substantial reliance by Presidential administrations, by the Department of Justice, by Senate committees, and by the media on ratings produced by the ABA and by similar organizations. Since the Eisenhower Administration, the ABA has enjoyed the privilege of "pre-clearing" the list of presumptive nominees put forth by the White House, and journalistic evidence suggests that dozens of candidates preliminary deemed "not qualified" by the ABA have had their candidacies derailed. Here, I have presented evidence that this rating process could be resulting in systematic bias against women and minorities. I have also presented evidence that these ratings are not particularly useful in terms of predicting long-term judicial performance. Taken together, they suggest that continuing to allow non-governmental organizations like the ABA to "pre-clear" nominees provides little benefit.

2.9 Appendix

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	−50.01*	−49.67*	47.95	47.67	58.63
	(9.16)	(11.00)	(68.42)	(68.51)	(74.80)
Age	0.06*	0.06*	0.06*	0.06*	0.06*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
African American	−0.81*	−0.99*	−0.98*	−0.94*	−1.18*
	(0.20)	(0.22)	(0.22)	(0.24)	(0.26)
Hispanic	−0.74*	−0.64*	−0.65*	−0.72*	−0.77*
	(0.23)	(0.27)	(0.27)	(0.30)	(0.36)
Female	−0.41*	−0.50*	−0.50*	−0.51*	−0.60*
	(0.16)	(0.18)	(0.18)	(0.19)	(0.21)
Republican	−0.04	−0.14	−0.04	−0.04	−0.48
	(0.11)	(0.23)	(0.36)	(0.36)	(0.43)
Top 25 Law School	0.31*	0.33*	0.34*	0.34*	0.13
	(0.11)	(0.12)	(0.12)	(0.12)	(0.15)
Private Law School	0.11	0.17	0.17	0.17	0.03
	(0.11)	(0.12)	(0.12)	(0.12)	(0.14)
Law Clerk	0.20	0.20	0.20	0.20	0.22
	(0.14)	(0.16)	(0.16)	(0.16)	(0.17)
Law Professor	−0.04	−0.02	−0.01	−0.01	−0.01
	(0.23)	(0.24)	(0.24)	(0.24)	(0.26)
Private Practice	0.51*	0.49*	0.52*	0.52*	0.50*
	(0.20)	(0.22)	(0.23)	(0.23)	(0.24)
US Attorney	−0.23	−0.25	−0.23	−0.23	−0.23
	(0.20)	(0.21)	(0.21)	(0.21)	(0.23)
Assistant.US.Attorney	0.63*	0.63*	0.63*	0.62*	0.66*
	(0.15)	(0.16)	(0.16)	(0.16)	(0.18)
Justice Department	0.13	0.16	0.15	0.15	−0.06
	(0.26)	(0.30)	(0.30)	(0.30)	(0.33)
Public Defender	0.20	0.50	0.54	0.55	0.38
	(0.28)	(0.32)	(0.32)	(0.32)	(0.35)
Federal Magistrate	0.79*	0.66*	0.64*	0.64*	0.72*
	(0.22)	(0.24)	(0.24)	(0.24)	(0.26)
Federal Bankruptcy	0.82	0.88	0.89	0.91	0.91
	(0.48)	(0.49)	(0.49)	(0.49)	(0.55)
State Judge	0.27*	0.27*	0.27*	0.27*	0.17
	(0.11)	(0.12)	(0.12)	(0.12)	(0.13)
Commission Year	0.02*	0.02*	−0.03	−0.03	−0.03
	(0.00)	(0.01)	(0.03)	(0.03)	(0.04)
ideology		0.04	−0.09	−0.08	0.47
		(0.32)	(0.33)	(0.33)	(0.45)
African American*Female				−0.21	−0.10
				(0.56)	(0.58)
Hispanic*Female				0.32	0.36
				(0.63)	(0.67)
Presidential Administration Dummies			✓	✓	✓
District Court Dummies					✓
N	1629	1421	1421	1421	1421

Standard errors in parentheses
* indicates significance at $p < 0.05$

Table 2.11: Logit regression results (judges nominated 1960-2002). Outcome is high ABA rating.

3

Example II: The Effect of Race in Judicial Appellate Review

In 1961, Illinois state judge James Parsons was at his summer home when he got a phone call that changed his life. The call was from President John F. Kennedy, and over the course of the call, Kennedy asked Parsons if he would accept a federal judgeship at the U.S. District Court for the Northern District of Illinois. As Parsons later recalled, "I said, 'As a former naval officer, aye, aye sir,' And he said, 'Carry on.'"

The significance of this conversation – an otherwise routine exchange between a President and a potential judicial nominee – was that James Parsons was black, and his investiture made him the first African American appointed to a lifetime federal judgeship in the U.S. District Courts. Parsons would go on to enjoy an illustrious career, eventually being appointed Chief Judge of his district and becoming the Seventh Circuit's representative to the Judicial Conference of the United States.

Thanks to jurists like Judge Parsons, numerous men and women of color now occupy roles in the upper echelons of the judiciary – not just in state and federal courts, but also in other countries and at the international level. And while social scientists have a good understanding of how individual characteristics such as race influence decision making (Kastellec, 2011), less well understood is how the legal system has incorporated these actors – that is, how the decisions rendered by minority and women judges have been evaluated by higher courts, whether they have been treated on equal footing, and how influential they have been. On the one hand, the increased appointment of women and judges of color serves to make the judiciary more reflective of the population it serves. On the other, if these judges are more likely to be overturned, then we must consider whether more needs to be done to achieve the goals of descriptive representation in the courts.

In this chapter, I offer the first study exploring how higher-court judges evaluate opinions written by African-American judges. Focusing on the U.S. federal courts, I leverage the fact that incoming cases are randomly assigned to judges within the same jurisdiction, which ensures that black and white judges on average hear similar sorts of cases. By then matching on measures of judge qualifications (including "quality" ratings assigned by the American Bar Association), professional and judicial experience, caliber of legal training, and partisanship, I find that cases decided by

African-American lower court judges are consistently overturned more often than cases authored by similar white judges. The effect is robust and varies little across legal issue, geographic region, or time. I find that this effect is particularly strong for cases authored by judges appointed by Democrat presidents.

The mechanism underlying these results is not straightforward. One possible explanation is that the difference is being driven by differences in ideological views (perhaps because black judges are more liberal). To test this, I examine whether black judges are more or less likely to be overruled by conservative higher-courts. I find that the difference between black and white judges in terms of reversal does not vary across more or less conservative higher courts. I therefore rule out the possibility that the difference between black and white judges is explained exclusively by ideological differences. A more likely possibility is that there are a variety of factors at play – including possible racial bias by higher courts.

This chapter proceeds as follows. Existing theories of judicial demographics and evaluation are discussed in Part 1. In Part 2, I discuss the data, which are 11,000 randomly selected cases taken from the U.S. Courts of Appeals from 1960-2002. Parts 3 and 4 discuss the methodology used, making particular note of the random assignment of cases to judges. Part 5 presents the results, while Parts 6, 7, and 8 discuss the causal mechanisms that could explain the observed patterns, focusing on (1) possible differences in ideology, (2) judicial qualifications, and (3) racial bias by higher courts. I conclude with a discussion of both the limitations and the implications of this research.

3.1 Judicial Demographics and Review by Higher Courts

Scholarship in judicial decision making has largely focused on how individual traits, including race, influence one's own opinions.[1] For example, in the criminal context, Scherer (2004), finds that black judges are more likely to accept black defendants' claims of police misconduct, while Welch, Combs and Gruhl (1988), Gottschall (1983), and Spohn (1990) find that black judges are more likely to be lenient with black defendants compared to white judges. In the civil rights context, Kastellec (2011) finds that a black presence on a three-judge federal appeals panel will make the panel more likely to support affirmative action programs than if it had no black judges serving; similarly, Pinello (2003) finds that black judges are more likely to side with LGBT claimants than white judges, and Martin and Pyle (1999) find that black judges are more likely to rule in a liberal direction in discrimination cases and gender-related cases. (On this last point, however, Segal (2000), finds evidence to the contrary.) A sizable number of studies have, however, found no difference between African-American and white judges across a variety of substantive legal areas (Walker and Barrow, 2009; Gottschall, 1983).

The literature thins significantly when it comes to how minority judges are evaluated or perceived by *other legal actors*. What limited insight we have comes from state-level analyses because, unlike federal judges, many state judges are elected directly by voters. This fact has led to institutional attempts to evaluate and quantify judicial performance in anticipation of judicial elections. These kinds of judicial evaluations have been implemented in 19 states and usually involve surveys of local attorneys about judicial

[1] Prior research here has analyzed the effect on decision making of judges' (1) political ideology and partisanship (Segal and Spaeth, 2002; Sunstein et al., 2006), (2) gender (Boyd, Epstein and Martin, 2010; Peresie, 2005; Smith Jr, 2005), and (3) professional experiences (Epstein, Knight and Martin, 2003; Sisk, Heise and Morriss, 1998; Brudney, Schiavoni and Merritt, 1999).

performance (Pelander, 1998; Wood, Lazos and Waters, 2010). These sorts of survey-based evaluations are not without their critics (Kearney, 1999; IAALS, 2008); Wood, Lazos and Waters (2010) have, for example, found that women and minority judges are awarded lower scores in the evaluation surveys even after controlling for objective measures of judicial qualifications, including the prestige of the law school attended, previous judicial experience, and reversal rates.

At the federal level, no study has looked at the comparative performance of minority or women judges, or at how often these judges are overturned by higher courts. To the contrary, federal judges are not evaluated formally, and there is an active debate on how best to institute accountability and evaluation procedures (Kourlis and Singer, 2008). Perhaps the only measure of judicial "quality" comes in the form of ratings assigned to judicial nominees by the American Bar Association (ABA). These ABA ratings are released to the public at time of nomination and have in the past proved controversial, and the administration of President George W. Bush withdrew its participation in the ABA rating process (although White House participation has since been re-instituted under the Obama administration). In addition, the ABA ratings have shown by some studies to be biased in favor of liberal nominees (Vining, Steigerwalt and Smelcher, 2009) or candidates with more experience (Haire, 2001), or to be completely unassociated with whether a judge will be more consistently upheld or overturned (de Rohan Barondes, 2009). Neither do the ABA scores assess in any way judicial performance *after* a judge's investiture – the issue that political actors and legal observers are most interested in. Nonetheless, ABA scores may assess attributes of judicial quality that objective criteria do not, and I discuss their potential role in greater depth below.

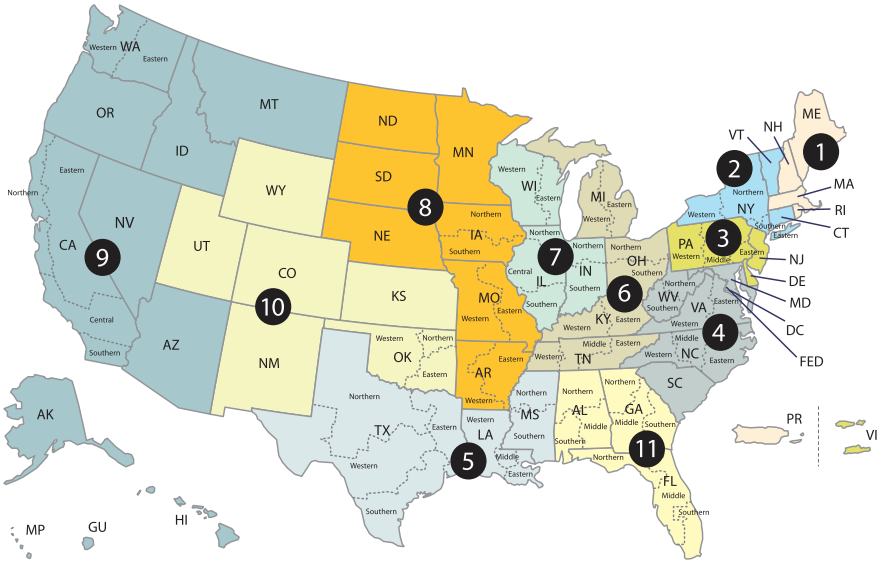

Figure 3.1: Numbered areas represent the boundaries of the U.S. Courts of Appeals, while dashed lines represent the boundaries of the smaller U.S. District Courts (Source: Federal Judicial Center).

3.2 Data

The data[2] for this analysis are 10,957 randomly selected cases (1960-2002) from the two lower tiers of the federal judiciary – the Courts of Appeals and the District Courts. The lower District Courts are the federal judiciary's workhorses, and, with 94 currently operating courts (organized geographically with at least one in each state), they hear the largest number of cases. By contrast, the middle tier, the U.S. Courts of Appeals, hears cases that litigants choose to appeal from lower court decisions. (Because losing litigants have the option of appealing, not all cases are appealed. This could introduce bias, a possibility I discuss below.) As such, the appeals courts are smaller and have

[2] All data, and accompanying replication R code, will be made publicly available at the conclusion of this project.

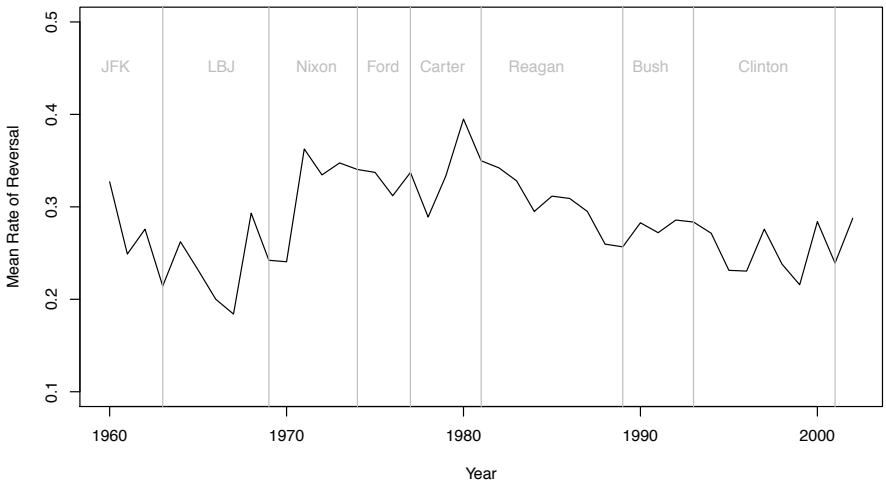

Figure 3.2: Mean reversal rates at the U.S. Courts of Appeals (for all judges), by year.

larger jurisdictions: there are 12 Courts of Appeals (known as "Circuits"), one for each of 11 geographical areas and one for the District of Columbia.[3] Importantly, district court judges decide cases alone, which makes it easier to determine the influence of a particular judge's race or ethnicity; by contrast, appeals judges almost always hear cases in panels of three. Also important is the fact that appeals judges have met most of their lower-court counterparts and will therefore be aware of the the race, gender, and basic demographics of these judges.[4]

To examine how ascriptive characteristics of lower-court judges could influence the outcomes of these 10,957 cases, I collected demographic information about each

[3] To give some context, approximately 310,364 cases were filed in U.S. District Courts in 2000; approximately 54,697 cases were filed the same year in the U.S. Courts of Appeals.

[4] This a key assumption, but one borne out by the fact that higher- and lower-court judges interact personally (by frequently having offices in the same building) and professionally (by routinely participating in judicial conferences and other professional organizations). Dropping the jurisdiction least likely to meet this assumption – the extremely large Ninth Circuit – did not meaningfully alter the results.

district court judge who served between 1960 and 2002. Much of this data came from a collection of biographical information compiled by the U.S. Federal Judiciary Center,[5] with additional information from a data base on judicial attributes collected by Zuk, Barrow and Gryski (2009). For each judge, I had data on his or her (1) race or ethnicity, (2) year of birth, (3) gender, (4) law school attended, and (5) geographic location. To gain purchase on judicial qualifications, I also collected data on each judge's professional experience, including whether he or she had worked as a U.S. Attorney, in the Solicitor General's office, as a law professor, or in private practice. Because more ideologically extreme candidates may be nominated when the Senate is on break, I noted whether the appointment was a "recess" appointment. Lastly, for partisanship, I recorded for each judge (1) the party of the appointing President, and (2) his or her judicial common space score (Boyd, 2011; Giles, Hettinger and Peppers, 2001; Epstein et al., 2007; Poole, 1998), which relies on the common space score of either the senior senator of the residing state or of the appointing President.

I also collected information basic information about the 10,957 cases, of which 537 were originally decided by black lower-court judges and 10,420 by white lower-court judges. Here, I used a database by Songer, Kuersten and Haire (2007) and, for each case, I have data on the lower court judge, the identity of the higher-court judges hearing the appeal, and whether the lower-court opinion was (1) upheld or (2) reversed (Figure 3.2). This is measured as a dichotomous variable, i.e., the case was upheld or it was not. (More sophisticated measurements of reversal, such as whether a case was reversed in part and affirmed in part, yielded similar substantive answers.) Other data include the year the case was decided and its substantive issue area.

[5] http://www.fjc.gov/history/home.nsf/page/judges.html.

3.3 Case Assignment and the Comparability of Judges

Key to the analysis is the fact that incoming cases in both jurisdictions – both in district courts *and* in courts of appeals – are assigned to judges on a random basis.[6] Although the randomization procedures are essentially ad hoc (certainly by experimental standards), the long standing practice of "random assignment of cases" makes it impermissible for federal judges to request to hear particular kinds of cases. Thus, in this context, black lower-court judges may never request to hear certain kinds of cases (for example, cases involving civil rights or black criminal defendants); likewise, possibly discriminatory white higher-court judges may never request to hear cases written by black lower-court judges. From a causal perspective, the two levels of randomization therefore work together to ensure that (1) on average, cases heard by black lower-court judges should be similar to those heard by white judges (i.e., there should be covariate balance in background case characteristics between cases heard by black judges versus those heard by white judges) and (2) appeals panels hearing cases written by black judges should on average be similar to appeals panels hearing cases written by white judges (that is, potentially biased judges cannot request to hear cases decided by lower-court black judges).[7]

[6] For example, based on personal phone calls to all Courts of Appeals offices, the randomization in the appeals courts currently works roughly in two ways. First, judges are may be randomly grouped together in three-judge panels and, second, cases are randomly assigned to the panels. Six circuits – the First, Fourth, Fifth, Sixth, Eighth, and Tenth Circuits – have a more formal mechanism in place. In these circuits, panels for each sitting are determined in advance. In the Fourth, Eighth, and Tenth Circuits this is done through the use of a computer program that achieves randomization within a certain set of constraints; assignment in the other circuits is done randomly by hand. Once the panels are set, a court official will randomly assign to the panels cases that are ready for review. In the Fourth, Eighth, and Tenth Circuits this is done through the use of a computer program that achieves complete randomization across panels, while in the First, Fifth, and Sixth Circuits, this random assignment is done "by hand."

[7] This also means, from a causal perspective, that the moment of "treatment" is the point at which a case is randomly assigned to a panel. Because of post-treatment bias issues, the timing of treatment is particularly delicate when dealing with race and ethnicity (Ho et al., 2007; Greiner and Rubin, 2010). Here, the treatment (authorship by a black versus white judge) is administered randomly at a specific point in time; we

Circuit	Judge Race	Criminal	Civ Rights	1st Am	Due Process	Labor	Econ
DC	Whites	34.3	9.7	4.0	5.6	5.3	37.1
	Blacks	27.5	10.1	3.7	9.2	2.8	36.7
6th	Whites	37.1	13.0	1.4	1.0	6.2	38.3
	Blacks	41.3	17.4	4.3	2.2	4.3	23.9
4th	Whites	32.7	11.2	3.2	1.5	3.5	44.4
	Blacks	24.4	17.8	6.7	2.2	8.9	40.0
2d	Whites	34.3	10.6	2.2	1.2	5.0	43.6
	Blacks	31.6	15.8	0.0	0.0	1.8	49.1
7th	Whites	34.4	13.4	2.4	0.9	4.6	41.8
	Blacks	31.0	17.2	3.4	1.7	5.2	34.5
5th	Whites	37.6	12.2	1.5	0.9	4.2	41.0
	Blacks	38.9	22.2	11.1	0.0	0.0	27.8

Table 3.1: Distribution of case issue area by race (1960-2002) for jurisdictions with highest % of black judges.

Nonetheless, it is possible that the randomization process used by the courts could break down, resulting in black and white judges within the same district (or appeals circuit) systematically hearing different kinds of cases.[8] To double check, I compare the cases written by (and eventually appealed from) black and white judges, the results for which are shown in Table 3.1 for the jurisdictions with the highest concentration of African American judges. (The distributions look similar in other circuits). For the most part, the cases heard by black and white judges are comparable, an intuition borne out by Table 3.2, which shows the results of a series of χ^2-square tests examining a null hypothesis of no relationship between the race of the lower-court judge and substantive case legal issue are, conditional on jurisdiction. Across all of the jurisdictions, I cannot reject the null hypothesis that there is no relationship between

can therefore control on those background variables that are clearly administered before cases are assigned to panels without inducing post-treatment bias. We also have a well-defined experimental analogy.

[8] A possible concern is judges in semi-retirement ("senior status" judges). Senior status judges in some jurisdictions elect to sit on judicial panels, thus raising the possibility that they could choose to hear certain kinds of cases. As the next discussion shows, the matching used left no senior-status judges remaining in the matched sample, thus mitigating this concern.

Circuit	p-value	Significant?
1st	0.73	No
2d	0.71	No
3d	0.54	No
4th	0.33	No
5th	0.11	No
6th	0.22	No
7th	0.43	No
8th	0.17	No
9th	0.51	No
10th	0.11	No
11th	0.36	No
DC	0.17	No

Table 3.2: χ^2 tests of difference between black and white judges' cases across legal issue area.

judge race and cases heard. This suggests that the random assignment of case assignments is – for purposes of this analysis, at least – fairly effective, and that black and white judges are hearing similar sorts of cases. I consider other possible sources of biases having to do with the appeals process below.

That the randomization appears to be effective (at least when it comes to the race of the lower-court judge) does not change the fact that African-American judges might differ from white judges across key demographic criteria. (See Table 3.3.) A much higher proportion of African-American district court judges have, for example, been appointed by Democrat presidents compared to white judges (74% compared to 44%) and a higher proportion of them have been women (18% compared to 10%). In addition, African-American judges are more likely to be former law professors, to have graduated from those law schools traditionally identified as the "Top 14" (or T14) or from private law schools,[9] to have served as Assistant U.S. Attorneys, and to have

[9] The fact that more African-American judges attended private law schools may be a consequence of segregation, as many public law schools (particularly in the South) were closed to African-American students

	Whites	African Americans
Average proportion of cases upheld on appeal	0.72	0.66
Average Age at Investiture	50.15	48.17
Proportion who are...		
Female	0.10	0.18
Appointed by Democrats	0.44	0.74
Attended Private Law School	0.53	0.73
Attended Top 14 Law School	0.32	0.35
Attended Top Tier Law School	0.80	0.84
Attended Howard Law School	0.00	0.14
Former Law Professor	0.05	0.14
Former US Attorneys	0.09	0.03
Assistant US Attorneys	0.16	0.23
Worked at Justice Department	0.04	0.08
Worked in Private Practice	0.94	0.77
Former U.S. Magistrate Judges	0.05	0.03
Former State Supreme Court Judges	0.04	0.04
Rated "Well Qualified" by ABA	0.46	0.30
Rated "Qualified" by ABA	0.40	0.65
N	1086	91

Table 3.3: Demographics of U.S. District Court Judges confirmed after 1960.

worked in the Justice Department in some capacity. African-American judges are also more likely to have attended Howard University Law School, with 14% of all black judges (compared to no white judges) calling Howard their alma mater. White judges, on the other hand, are more likely to have experience in private practice (94% compared to 77%) and to have served as U.S. Attorneys (as opposed to *assistant* U.S. Attorneys, 9% compared to 3%). White judges are also more likely to have cases be upheld on appeal: a white judge will on average have 72% of his or her cases upheld on appeal compared to 67% for black judges.

In addition, case randomization at the district level does not stop litigants from

into the 1950s.

	1st	2d	3d	4th	5th	6th	7th	8th	9th	10th	11th	DC
White	0.64	0.76	0.78	0.85	0.83	0.86	0.79	0.82	0.75	0.78	0.85	0.57
Black	0.03	0.06	0.06	0.08	0.04	0.09	0.07	0.06	0.07	0.03	0.11	0.22
Other	0.32	0.18	0.16	0.07	0.14	0.05	0.14	0.12	0.19	0.19	0.04	0.22
N	59	142	140	113	213	130	99	104	208	78	55	37

Table 3.4: Racial breakdown of district court judges by appeals circuit (for all judges confirmed after 1960). Judges in the "other" category include Hispanics, Asian Americans, and Native Americans. (The First Circuit, which includes Puerto Rico, therefore has a high percentage of judges in the "Other" category.)

bringing cases in different jurisdictions. Different types of cases originate in different parts of the country (e.g., urban versus agricultural areas, or coastal states versus rural states), and each circuit has different proportions of African-American judges (see Table 3.4), different numbers of Republican and Democrat appointees, or different norms about reversal. If any serious imbalance exists among background case characteristics, it could be due to forum shopping by litigants. I address this by including the U.S. district in which the case originated as a key variable in the analysis.

3.4 Methodology

Because black and white judges differ substantially in their age, previous employment, partisanship, and geographic dispersion, and because different cases arise in different jurisdictions, simple comparisons between the black and white judges provide limited insight. To account for differences, I use matching (Ho et al., 2007). Matching operates here by comparing cases written by judges who are identical across key characteristics. Thus, a lower-court opinion written by black judge sitting in the Eastern District of Louisiana who graduated from a second-tier law school with previous experience working as a state supreme court judge will be compared to a lower-court opinion written by a white judge also from the Eastern District of Louisiana with an exactly

similar profile.

This approach offers several advantages. First, matching is an effective pre-processing step that reduces dependence on statistical modeling assumptions (Ho et al., 2007). Second, and relatedly, matching effectively tests all possible ways that variables could interact with each other. We may, for example, think that black judges will be more likely to be overruled in southern circuits, or that black judges with top tier law school degrees may be less likely to be overruled. By pruning the data, matching resolves this problem and isolates the effect of a judge being African American, regardless of the possible ways that other variables may be affecting one another. To implement the matching, I use coarsened exact matching (Iacus, King and Porro, 2011, 2009), which allows exact matching on key variables and coarsening and then matching approximately on the three variables that are continuous (discussed below). Coarsened exact matching has the advantage of allowing for this approximation to be as close as needed to remove biases. I also have the advantage of matching exactly – the best form of matching – on a large portion of the variables measuring judicial qualifications.[10]

Once cases written by these judges were matched and other cases pruned, I took the difference in means in how often cases written by black judges were upheld versus those written by white judges.[11] Because lower-court judges may write multiple opinions over the course of their careers, I included judge-specific random effects, a modeling specification that accounts for the increased dependence between these observations. (The inclusion of random effects does not meaningfully alter the results.)

It should be noted that matching has some drawbacks, including the fact that many

[10] Using different matching estimators (nearest neighbor matching and propensity score matching) yielded similar substantive results, as did estimating the effect without discarding any "treated" units (i.e., cases written by black judges). I present the results from coarsened exact matching, as it bounds the maximal amount of imbalance through the choice of coarsenings (Iacus, King and Porro, 2011, 2009).

[11] I obtain this via a simple linear regression. Since both the independent and dependent variables are dichotomous, this imposes no functional form assumptions on the data.

observations will be dropped. For the core results presented, this is not a problem: sufficient observations remain after matching to make statistically significant inferences, and the matched sample by no means an anomalous subset of the entire universe of judges. As discussed below, however, we may be interested in estimating the effect over not just the subset of the population for which there is good overlap (e.g., the matched sample), but also over the full population of interest (e.g., all judges). We may also be interested in how the effect of having a black judge varies over certain population subsets – including across across different partisan configurations, different legal issues, or across different geographic jurisdictions; these may all have implications for the causal mechanism(s) behind the results. Thus, I at times fit mixed-effect models that allow the effect of race to vary, controlling for the same characteristics used in matching.[12] Because coefficients obtained using a logit link function can be difficult to interpret, I present predicted probabilities throughout.

Judge Attributes. At all times I match on, or control for, key personal characteristics of lower-court judges, including whether a judge (1) was male or female, (2) had ever served as a United States attorney or as an assistant United States attorney, (3) had served as any other kind of Department of Justice attorney or Congressional counsel, (4) had worked in the Solicitor General's Office (as the Solicitor General, or as a Deputy or Assistant Solicitor General), (5) had ever served as a state judge (either as a state supreme court or state lower court judge judge), (6) had ever been a former federal judge (e.g., magistrate, territorial, or bankruptcy judge), (7) had worked as a full-time law professor, or (8) had experience as an attorney in private practice. I also match on whether the judge attended a top tier law school[13] and

[12] I do so using the R package `lme4` (Bates, Maechler and Bolker, 2011).

[13] Because many of the judges in the data set attended law school at a time when (a) segregation was still practiced and (b) law school rankings were not assigned, I use a flexible definition of "top tier." Specifically,

	Whites		African Americans	
	All	Matched	All	Matched
Average Age at Investiture	50.15	49.55	48.17	47.75
Proportion who are...				
Female	0.10	0.04	0.18	0.04
Appointed by Democrat Presidents	0.44	0.75	0.74	0.75
Attended Private Law School	0.53	0.63	0.73	0.63
Attended Top Tier Law School	0.80	0.92	0.84	0.92
Former Law Professors	0.05	0	0.14	0
Former US Attorneys	0.09	0	0.04	0
Worked at Solicitor General's Office	0.01	0	0	0
Worked in Private Practice	0.94	0.96	0.77	0.96
Former U.S. Magistrate Judges	0.05	0	0.03	0
Former U.S. Bankruptcy Judges	0.01	0	0.04	0
Former State Supreme Court Judges	0.04	0	0.04	0
Former State Lower Court Judges	0.04	0.13	0.04	0.13
N	1086	67	91	24

Table 3.5: Demographics of matched district court judges compared to the entire population of district court judges.

whether the judge was a recess appointment. Finally, although controversial and alleged by some to be biased, the possibility exists that qualification ratings issued by the American Bar Association at time of nomination assess attributes of judicial acumen that objective criteria do not. I therefore match exactly on one of four possible ABA qualification rating: "Exceptionally Well Qualified," "Well Qualified," "Qualified," and "Not Qualified." (The "Exceptionally Well Qualified" rating was removed in June of 1989.) I discuss the role played by ABA cores in greater depth below.

I include those law schools now found in the U.S. News & World Report Top 100 law schools. I also include in this group Howard University Law School, the school of choice for many African Americans through the 1950s. (Dropping the Howard-educated judges and other graduates from historically black law schools from the analysis – perhaps on the assertion that there are no good matches for them – had the effect only of reducing the number of observations and introducing more uncertainty into the final estimate: the direction and rough magnitude of the effect were not affected.) I also include a dummy variable for whether the law school was public or private to account for possible segregation by public law schools through the 1950s.

For the matching, I coarsen three key variables: (1) year of birth, (2) years of U.S. District Court experience, and (3) political ideology. When taken together with the year the case was argued (discussed below), year of birth is an effective proxy for age. Because it is extremely difficult to match exactly on year of birth without losing much of the data, however, I coarsen this variable into 20-year cohorts. Years on the bench (i.e., federal judgeship experience) is also a key measure: we may think that judges with more experience will be less likely to be overturned. I coarsen this variable to compare judges with 0-5, 6-15, 16-30, 31-45, 46-60, and 60+ years of experience at the time the case was argued. This has the effect of comparing only those judges with similar federal district court experience to one another. Finally, I coarsen slightly the judicial common score for each judge (Boyd, 2011), while also including the party affiliation of the appointing President. Black and white judges differ in their political ideologies (see Figure 3.3), and I discuss the possible role of ideology in greater depth below.

A summary of lower-court judge characteristics post-matching is given by Table 3.5. This matched sample of judges is, as expected, slightly different than the original pre-matched sample (Table 3.3) but by no means anomalous. None of the matched judges had experience working as federal magistrate or bankruptcy judges, as law professors, or as U.S. Attorneys, a testament to the small number of such individuals in the population of judges at large. In addition, the matched sample has a greater proportion of male judges who attended a "Top Tier" law school, whose careers were spent in private practice, and who were appointed by Democrat Presidents. Lastly, the average years of experience for the post-matched sample, 8.7 for whites and 10.4 for blacks, demonstrates the scarcity of senior judges (who must have over 15 years of experience to take the status). This mitigates the concern that senior status judges, who have the option of participating in certain panels, are substantially disrupting the case

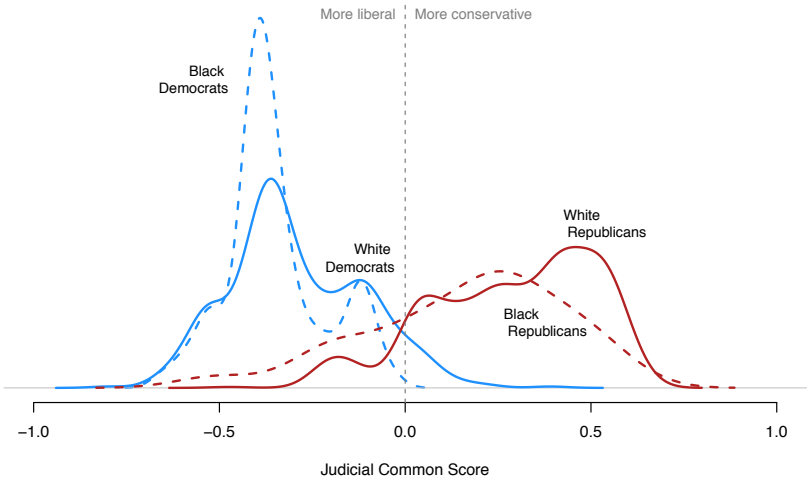

Figure 3.3: Distribution of Judicial Common Scores for black and white U.S. district court judges (confirmed after 1960), disaggregated by race and party of appointing president.

randomization and biasing the results.

Case and Court Attributes. In addition to these individual characteristics, I match on, or control for, various court and case attributes. First, to account for the possible influence of geography and forum shopping, I include the exact U.S. District in which the case originated. Second, I include the year the case was decided. I do so by comparing cases decided in twenty- or ten-year intervals: 1961-1980, 1981-1991, and 1992-2002.

It is also possible to match on, or control for, additional case factors, including the (1) party composition and (2) racial composition of the three-judge panel hearing the case's appeal. Conditional on a case arising out of the same jurisdiction and in the time frame, however, cases have the same approximate probability of being assigned to comparable appeals panels. Thus, much like a randomized medical experiment, once

we control for (match on) jurisdiction and time frame, black and white judges will be reviewed by similar appeals panels (both in terms of partisanship and in terms of racial composition). Nonetheless, it is possible that the randomization might not be working cleanly, and I checked this by matching on the party composition and the racial composition of the appeals panel. Although I do not include those results here, they are comparable in all respects to the ones presented here – which provides additional evidence that case randomization is working as expected, at least in this context. I do explore whether the effect varies across different higher-court partisan compositions (e.g., majority Republican versus majority Democrat) and racial compositions by including the results from interacted mixed-effects models (discussed below).

3.5 Results

I begin by presenting the pre- and post-matching results from 1960-2002. (See Figure 3.4.) The horizontal axis here, and in subsequent plots, represents the effect of having a black lower court judge on the probability that a case will be upheld by a higher court on appeal. For purposes of interpretation, a positive effect means that cases written by black judges are *more likely* than cases written by white judges to be upheld, while a negative effect means that cases written by black judges are *less likely* than cases written by white judges to be upheld. For example, a point estimate of -0.25 would mean that cases written by black judges are being overturned at a rate 25 percentage points exceeding those written by white judges.

Before matching, as the top line of Figure 3.4 demonstrates, a case having an African-American judge as its author is approximately 6% more likely to be overturned on appeal than one written by a white judge. The effect is significant (with a p-value of less than 0.05), meaning that we can rule out a non-existent relationship between black

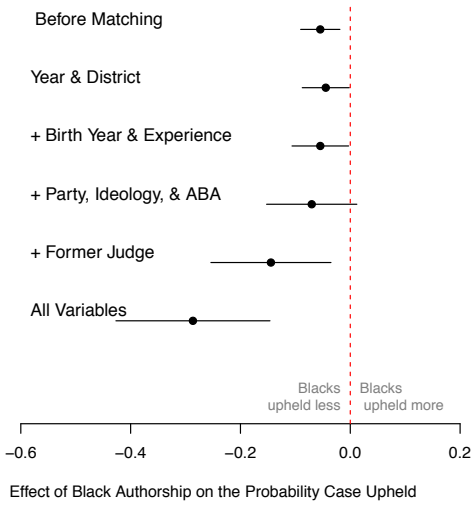

Figure 3.4: Effect of black authorship on probability case upheld at court of appeals. Estimates obtained by taking a difference-in-means (with judge-specific random effects) before matching, incrementally matching on additional variables. Solid dots are point estimates; lines represent 90% confidence intervals.

lower-judge authorship and its likelihood of being reversed. Once we include additional variables into the matching – including year the case was decided and U.S. District, age (via birth year) and professional characteristics, judicial experience, ABA qualification ratings, and educational background, the difference only widens. Once we include all variables into the matching, we see an approximate 20% difference between black and white judges; that is, a case written by a black judge is up to 20% more likely than one written by a white judge to be reversed on appeal. This difference is again statistically significant, with a p-value of less than 0.05.

Important to note is that a large majority (nearly 75%) of African-American judges are Democrat appointees. To further shed light on the possible relationship between partisanship and reversal, I calculate the effect separately for Democrat-appointed and

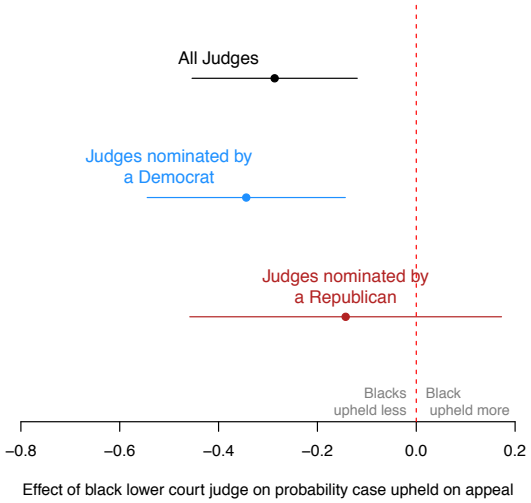

Figure 3.5: Effect of black authorship on probability case upheld at court of appeals (calculated separate by party of appointing president). Estimates obtained by taking a difference-in-means after matching (with judge-specific random effects). Solid dots are point estimates; lines represent 95% confidence intervals.

Republican-appointed judges. When I examine only opinions authored by Democrat-appointed judges (Figure 3.5), a case written by a black judge is approximately 30% more likely to be reversed than one authored by a white judge. Despite a decreased sample size, the effect is again significant. On the other hand, once we limit the sample to look only at Republican appointed judges, the effect of a case being written by a black lower court judge is *not* statistically significant. (Similar non-results for Republican-appointees are obtained from an interacted fixed-effects regression – i.e., this is not an artifact of matching pruning too many observations.) Thus, the effect of black authorship appears to be one driven primarily by differences between black and white Democrats.

3.6 Differences in Beliefs or Ideology

A possible explanation for the difference between black and white judges is that African-American judges have different legal philosophies than white judges, and that these differences cause black judges to be overturned at higher rates. For example, black judges might have a different understanding of the role of the law in resolving certain disputes or they may hold different political attitudes. Here, I consider two possibilities.

Black Judges Vote Differently on Certain Issues. As suggested by the judicial politics literature, it is possible that black judges vote differently than white judges, but that they do so only with regard to cases having a significant racial, ethnic, or civil rights dimension. This could include substantive issue areas involving criminal law and procedure (Scherer, 2004) or affirmative action and civil rights (Kastellec, 2011). Accordingly, we may expect black judges to be overruled most frequently in criminal law and civil rights, the two areas their views might differ the most from whites. Or, we may expect that black judges may be upheld more in these areas, with appeals judges acting differentially to the opinions of black judges on racially sensitive cases. In either scenario, we would expect that the "black judges effect" would vary significantly between possibly racialized areas (e.g., criminal law, civil rights law) and others.

To test this possibility, I include in the final post-matching model random effects and varying coefficients for several key issue areas, including economic activity, labor relations, civil rights, and criminal law. For the "issue area" explanation to hold sway, we would expect the effect of black authorship to differ across the two areas (civil rights

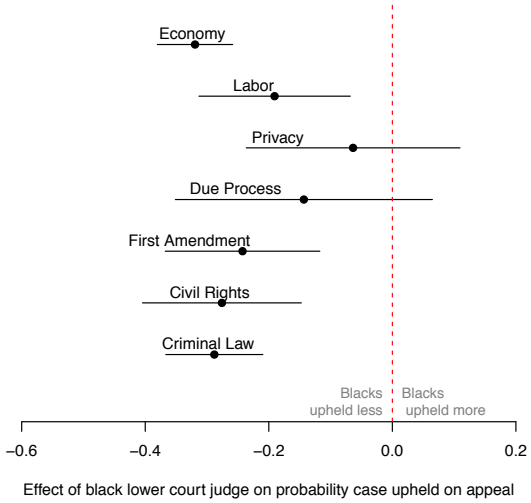

Figure 3.6: Effect of black authorship on probability case upheld at court of appeals by legal issue area. Estimates obtained via mixed-effect logit regression, controlling for (1) personal characteristics, (2) judicial experience and ABA ratings, and (3) case attributes. Judge-specific random effects included. Solid dots are point estimates; lines represent 95% confidence intervals.

and criminal law) identified by the judicial politics literature as being particularly racially salient. The results (Figure 3.6) show, however, that the effect of black authorship on a case's probability of being upheld actually varies little by issue area: indeed, across all the issue areas, cases authored by black lower-court judges are between 10% and 30% more likely to be overturned than those authored by whites. (Cases involving privacy and due process concerns fall just shy of statistical significance, more evidence of the small number of such cases than anything else.) Similar non-results are obtained from a fixed-effects regression that controls for the same characteristics included in the matching, while also interacting the race of the lower court judge with legal issue area: at no point are there statistically distinguishable differences in how the "black judges effect" varies across different substantive legal

issues. Thus, the data provide no evidence for the proposition that black judges are being overturned at greater or lesser rates (compared to white judges) within different legal categorizations. The "black judges effect" therefore appears unlikely to be driven exclusively by black and white lower-court judges deciding certain kinds of cases differently.

Ideologically Extreme Black Judges. Another plausible explanation is that black judges are more ideologically extreme across the board. On this point, some literature (Asmussen, 2011) suggests that Presidents who appoint "nontraditional" (e.g., minority and women) candidates take the opportunity to appoint more ideologically extreme individuals than they would otherwise. For African Americans, this practice would have the effect of introducing to the bench more ideologically extreme black candidates, who would then be overruled more by moderate appeals panels across all kinds of legal issue areas. This might particularly be the case among black Democrats, who are not only more numerous than black Republicans but, as a whole, more liberal than their white counterparts. (See Figure 3.3). The argument seems to hold less for black Republicans (again, see Figure 3.3), but making firm inferences about the pool of black Republicans is risky given its small size.

It is worth noting that the analysis at all times controls for partisanship of the appointing President, as well as judges' ideology in the form of their judicial common scores (Boyd, 2011), but a left-leaning difference in ideology could initially be mismeasured or develop over time.[14] To test this theory, I explore the possible interaction between the race of the lower court judge and the judicial common score *of*

[14] To give some context, the judicial common scores use either (1) the common score of the senior senator form the judge's state, or, if that senator is of the opposing party as the appointing president, (2) the common score of the president. (Boyd, 2011)

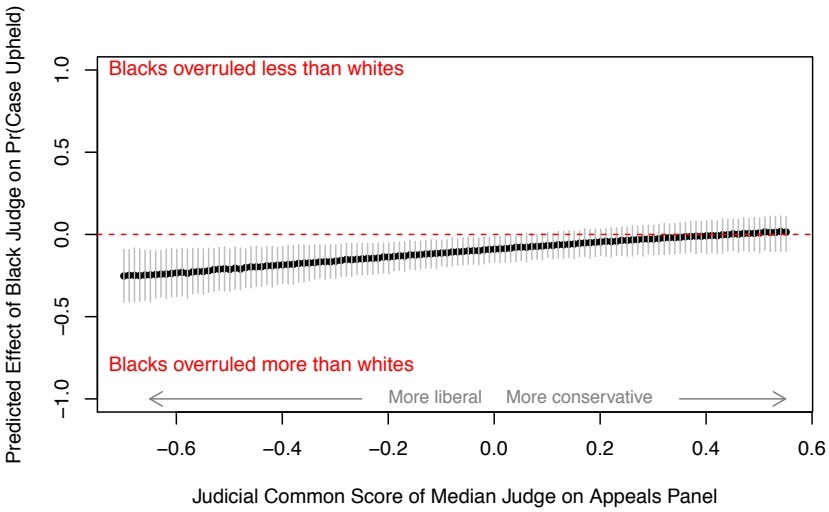

Figure 3.7: Predicted effect of black Democrat lower-court judge (as opposed to a white Democrat lower-court) on probability case upheld at court of appeals. Estimates obtained via mixed-effect logit regression, controlling for (1) personal characteristics, (2) judicial experience and ABA ratings, and (3) case attributes.on the probability that a case is upheld at the Court of Appeals. Judge-specific random effects included. Vertical lines represent 95% confidence intervals. Despite an apparent decrease in the effect as panels become more conservative, differences between liberal and conservative panels never reach statistically significance.

the median judge on the three-judge panel hearing the appeal. If black Democrats are more liberal than white Democrats, then we would expect an interactive relationship between the race of the lower court judge and the ideology of the appeals panel (as reflected by the median judge). That is, we would expect that conservative appeals panels would overturn black judges more than white judges at greater rates than would liberal panels.

In Figure 3.7, I present the results showing how the "black judge" effect varies according to the judicial common score of the median judge on the three-judge panel hearing the appeal. (The predicted probabilities here are calculated after fitting a fixed-effect logit regression, with race of the lower court judge interacted with the

judicial common score of the median judge on the appeals panel. I include the same controls as I did with the matching.) As Figure 3.7 demonstrates, the effect does not vary significantly: at no point are there statistically distinguishable differences in how the effect varies across different median judge judicial common scores. In addition, if anything, more conservative appeals panels appear *less* likely to overturn black judges more than white judges (although differences between liberal and conservative appeals panels in this regard are never statistically significant). In effect, we cannot rule out that the relationship of race on reversal in no way varies with the ideology of the median judge on the appeals panel. Thus, the partisanship of the appeals panel appears to have no mitigating or strengthening effect on the racial difference, and conservative and liberal appeals panels roughly are treating black judges the same (i.e., overruling them more than whites). This finding provides support against the theory that the effect is being driven by black democrats simply being more liberal.

3.7 Differences in qualifications or "quality"

An alternate explanation for these results is that African-American judges bring to the bench different qualifications and perhaps produce opinions that differ from those produced by white judges in terms of "quality" or legal acumen.[15] Any difference in such "quality" would have to persist despite comparing individuals with very similar pedigrees in education and experience (not impossible, but a somewhat high bar). For example, I control here for quality of legal education, including whether the judge attended a "top tier" school and whether the law school was public versus private. The

[15] Here, I use quotations purposefully, as there is little agreement on what constitutes a high-quality opinion and whether "quality" necessarily translates into greater or lesser rates of reversal: after all, "high-quality" yet ideologically extreme judges would still be reversed more. It is, however, agreed that there are some judges write with more legal precision and dexterity, while others are more frequently mistaken in how they apply basic legal principles; both could influence rates of reversal. Accordingly, it is on these pragmatic differences that I attempt to I focus this discussion.

definition is in fact loose because no rankings were used at the time these judges attended law school. Neither was it possible for certain black judges to attend law schools that were segregated until the 1950s (e.g., the University of Texas); neither did any white judges attend Howard University Law School, by far the most popular legal alma mater of black judges. Along these lines, a related argument is that white and black judges with similar experiences (both educational and professional) still enjoy different levels of success. Sander (2004), for example, finds that black lawyers who attended the same law schools as white lawyers were more likely to graduate at the bottom of the class and also had overall lower bar passage rates. (See, however, Ho (2005) for a rebuttal.) Controlling for attributes like law school or initial professional experience might therefore mask certain inequalities that could potentially explain differences in reversal rates.

ABA ratings and Judicial "Quality." To further disentangle the potentially fraught relationship between judge race and "quality," I examine more closely the American Bar Association scores awarded to each judge at time of nomination; these scores specifically purport to capture non-quantitative "quality criteria" such as "intellectual capacity, judgment, writing and analytical abilities, knowledge of the law, and breadth of professional experience" (American Bar Association, 2009). It is likely that the ABA scores take into account those attributes that are masked by the other variables, including class rank, the quality of law school coursework, bar exam passage, law review membership, etc. (However, the ABA's exact process is not disclosed and, in fact, the ABA "strictly maintains the confidentiality of its internal evaluation materials and reports" (American Bar Association, 2009).)

Table 3.6 provides the distribution of ABA qualification ratings for black and white

	Ex. Well Qualified	Well Qualified	Qualified	Not Qualified	*N*
Whites	0.04	0.51	0.44	0.01	981
Blacks	0	0.31	0.67	0.02	88

Table 3.6: ABA Ratings for U.S. district court judges confirmed after 1960. The "Exceptionally Well Qualified" category was discontinued in June of 1989; no black judges ever received this score.

judges confirmed after 1960: here, black judges tend to receive lower ratings than do white judges. One way to understand these ratings is by assuming that the ABA is an impartial (i.e., non-racially biased) and fairly accurate assessor of judicial quality. Under this understanding, we would expect that matching exactly on judges' ABA scores would account roughly for differences in legal acumen and dexterity between black and white judges: that is, conditional on having the same ABA rating, we should see no substantial differences in reversal rates between black and white judges. The results show, however, that cases written by black judges with the same ABA scores as whites are up to 20% more likely to be reversed than cases written by white judges. Thus, if we accept that the ABA is a accurate assessor of those subjective components of judicial quality or acumen,[16] then, by extension, the results presented above must be due to factors other than differences in judicial quality (i.e., possible racial bias, differences in legal and political views, etc.).

A more plausible scenario is that ABA ratings are (1) not perfect assessors of judicial quality and/or (2) are perhaps themselves related to the race of the nominee. Here, I consider two possibilities:

1. Perhaps because of racial bias, the ABA could be biased in favor of whites. In practice, this would mean that similarly situated whites are more likely to get

[16] de Rohan Barondes (2009) finds that ABA quality ratings are not predictive of whether a judge will be more or less overturned. Here, cases authored by black judges are overturned more frequently than cases authored by whites, regardless of whether ABA ratings are controlled for.

higher ABA ratings – i.e., within the same ABA ratings cohort, whites on average are of poorer quality than blacks.

2. For contrasting reasons, the ABA could be biased in favor of blacks. This would mean that, within the same ABA ratings cohort, blacks on average are of poorer quality than whites.

Assuming no subsequent racial biases by appeals panels, the two theories have different observable implications. The ABA being biased in favor of whites would mean that that controlling for ABA scores would bias the results in favor of whites being overturned more. That is, because whites are more poorly qualified than blacks with identical ABA ratings, comparing judges with identical ABA ratings should mean that whites should be overturned more. On the other hand, the ABA being biased in favor of blacks (an "affirmative action" story) would mean the opposite: because blacks are more poorly qualified than identically ABA-rated whites, they should be overruled more. Importantly, this would have the effect of explaining the results seen here based on "quality" (or "affirmative action") grounds.

Because the second of these two narratives could provide a theoretical explanation for the results, I examine more closely the possibility that the ABA could be more favorable to African Americans (i.e., that there could be an "affirmative action" explanation). I do so using a similar methodology to that described above (albeit in a non-randomized setting),[17] matching black and white judges across the slew of "objective" quality measures, which included many of the attributes matched on before: (1) gender, (2) party of appointing president, (2) year of birth, (3) circuit, (4)

[17] The lack of randomization means that making a causal claim is more difficult: the "treatment" (a judge's race) is assigned at birth, rendering the host of attributes I am matching on post treatment. My interest is, however, not the lifetime impact of a judge being black; rather, it is the way that the ABA responds to similarly situated black and white judges (Greiner and Rubin, 2010; Sen and Wasow, 2011).

past state court experience, (5) past U.S. attorney experience, (4) past solicitor general office experience, (5) past federal magistrate or territorial judge experience, (6) past law professor or private practice experience, (7) top-100 law school rank or a private law school, and (8) nomination year. I do not include attributes that could be affected by the ABA scores themselves (as this would introduce post-treatment bias). For example, a President's decision to push a nominee forward during Senate recess is an attribute that could be directly influenced directly by a judge's poor ABA rating.

Once I matched on these objective qualifications, I take a simple difference in means to gauge the difference between black and white judges in terms of the ABA ratings received. I use all four possible ABA ratings, from "Exceptionally Qualified" to "Not Qualified"; dichotomizing the ABA scores into (1) highly qualified and (2) either qualified or not qualified (as is commonly done in the judicial politics literature) did not change the results, which are presented in Figure 3.8. As before, the horizontal axis here represents the relationship between a district court nominee being black (as opposed to white) on the probability of receiving a particular ABA qualification rating.

For purposes of interpretation, a positive effect means that black nominees are on average *more likely* than matched white nominees to receive that rating, while a negative effect means black judges are less likely to receive that rating. For example, the estimate of −0.24 for "Well Qualified" means that black judges are over 20% less likely to receive this rating (currently the best rating) than are whites. Black judges were also less likely to receive the highest"Exceptionally Well Qualified" rating, although because the rating is no longer dispensed (and no black judges received this category), the point is somewhat moot. By contrast, black judges appear more likely to receive the second lowest rating ("Qualified") as well as more likely to receive the lowest rating ("Not Qualified") although the former falls shy of statistical significance.

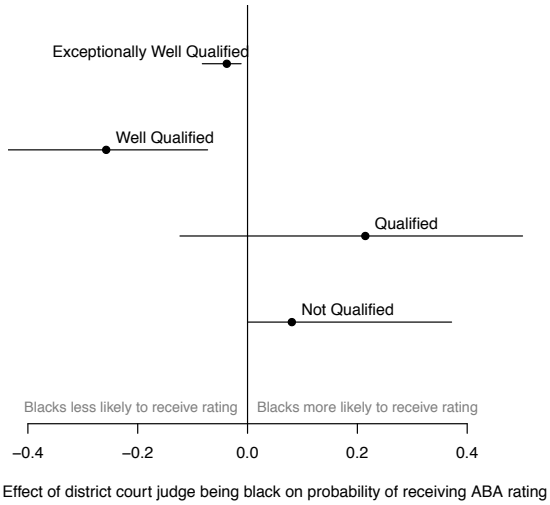

Figure 3.8: Relationship between judge race and probability of receiving certain ABA ratings. Estimates obtained by taking a difference-in-means after matching. Solid dots are point estimates; lines represent 95% confidence intervals.

These results allow us to rule out one possible explanation: that the ABA quality ratings are biased *toward* African Americans. The ABA scores being biased in favor of African Americans would have meant that controlling for ABA scores would in effect be comparing black judges of a poorer "quality" than white judges – a theoretical and empirical explanation for the results seem here. However, the fact that the ABA scores appear, if anything, tilted toward white judicial candidates rules out this kind of "affirmative action" explanation. (Indeed, the fact that ABA scores appear biased against African Americans raises the possibility that the results presented here are conservative.) The possibility exists that ABA scores are entirety unrelated to judicial quality, and, indeed, there is debate about the utility of these scores. Nonetheless, unless the ABA qualification system results in scores entirely orthogonal to some subjective sense of judicial "quality" (unlikely), then we can rule out that the results

presented here are being driven exclusively by subjective or unmeasured differences in judicial "quality."

Quality and Partisanship. Another consideration counseling against an exclusive "quality" explanation is that the effect of black-judge authorship on reversal varies depending according to the party of the appointing President, with the effect being driven by Democrat appointments. That is, a broad difference in the quality of black versus white candidates would have an observable implication in that all black judges (not just Democrat-appointed ones) would be more likely to be overturned. A distinct possibility is that black Democrats are on average less qualified than black Republicans, but the data do not bear this out – at least not when it comes to objective measures (Table 3.7). Indeed, black Democrats and black Republicans appear relatively well balanced – approximately equal proportions were U.S. attorneys (or Assistant U.S. Attorneys) or had previous state court experience. An equal proportion attended a top tier law school (including the modal school, Harvard Law School). Any imbalances between these two groups points to black Democrats, if anything, being more highly qualified: a higher proportion of black Democrats had experience working as law professors, had experience working in private practice, and attended a private law school. A mechanism reliant exclusively on a "quality" interpretation actually does little to explain the discrepancies in outcomes between these two comparable groups.

3.8 Racial Bias

Another possibility is that appeals judges are biased in their evaluation of opinions authored by lower-court African-American judges. On the one hand, this explanation has the deepest and most troubling normative implication as challenges the fairness

	Black Democrats	Black Republicans
Average Age at Investiture	47.98	48.67
Female	0.19	0.12
Attended Private Law School	0.77	0.59
Attended Top Tier Law School	0.16	0.17
Attended Harvard Law School	0.07	0.08
Former Law Professors	0.19	0
Former Assistant US Attorneys	0.22	0.25
Worked at Justice Department	0.09	0.04
Worked in Private Practice	0.84	0.58
Former U.S. Magistrate Judges	0.03	0.04
Former State Supreme Court Judges	0.04	0.04
N	67	24

Table 3.7: Demographics of black U.S. District Court Judges confirmed after 1960.

and race neutrality of the judiciary. On the other hand, such a finding would perhaps be unsurprising, as studies have teased out salient biases against racial minorities in prominent economic, social, and political settings, including in employment (Bertrand and Mullainathan, 2004), in housing (Yinger, 1986), and in academia (Ginther et al., 2011).)

Discrimination in an observational context such as this is extremely difficult to test; no variable measuring "racial bias" exists and, short of a fully randomized experiment, the case for racial bias is circumstantial. A possible test of this theory is to check whether the presence of a black judge on an appeals panel attenuates the effect. (Of the cases decided between 1960 and 2002, 88% had no black judges on the 3-judge appeals panel, 11% had one black judge, and just under 1% had two.[18] No cases in the data were heard by an all-black 3-judge panel.) Here, my supposition is that having at least one black judge on the appeals panel might help mitigate the black judges effect. This

[18] Here, I am aware that there is significant missingness in these data due to the presence on appeals panels of visiting judges, about whom I did not have personal demographic data. It seems unlikely that the missingness would bias the results seen here.

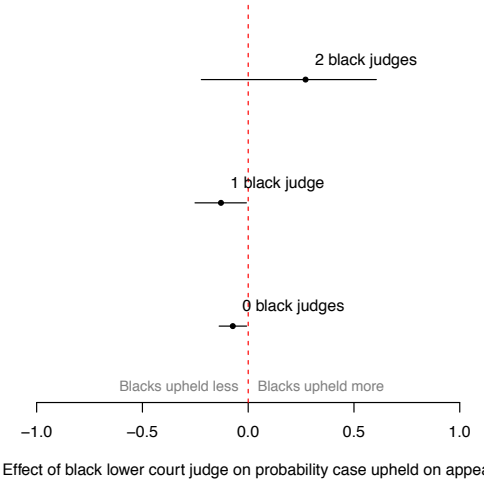

Figure 3.9: Effect of black authorship on probability case upheld, by racial composition of the appeals panel. Estimates obtained via mixed-effect logit regression, controlling for (1) personal characteristics, (2) judicial experience and ABA ratings, and (3) case attributes. Judge-specific random effects included. Solid dots are point estimates; lines represent 95% confidence intervals.

could happen as white judges become more sensitive to any possible discriminatory tendencies (on this point, Kastellec (2011) provides evidence that having a black judge on an appeals panel will change the way that panel votes), or as black judges raise concerns about bias. To test this theory, I evaluated how the black judge effect varied across different racial panel compositions – zero, one, or two black judges on the three-judge higher-court panels. The scarcity of black judges at the appeals level meant that matching was an inefficient methodology; I therefore use a mixed-effect model, controlling for the same judge and case characteristics, with an interaction between race of the lower court judge and the number of black judges hearing the appeal.

Results from this analysis are presented in Figure 3.9. Because of the exceedingly

low numbers of black judges on appeals courts, and because of the fact that these judges very rarely sit together, we cannot statistically distinguish panels heard by zero, one, or two black judges on the appeals. (Indeed, of the 24 instances where two or more black judges were on the appeals panel, only 5 of these were instances where the lower court judge was black.) Nonetheless, the trend is obvious – having 0 or 1 black judge is associated with a constant size of the "black judges effect." However, increasing the number of black judges on the appeals panel to two (i.e., to instances where black judges represent the majority), the effect disappears and not only loses significance, but also moves in a more positive direction. I also note, however, that this doesn't rule out the possibility that black appeals judges are just more liberal than white judges, an implication that would also explain the results on more ideological grounds.

3.9 Additional Mechanisms

Appeals Bias. A possible complication is that black and white reversal rates could be rooted in litigant choice on whether to appeal. Because appeals are taken at the behest of the losing party at the lower level, not all cases are appealed and documented in the data. This presents a structural missing data problem: ideally, we would like to know the outcome of all cases, regardless of appeal, but appeals courts cannot and do not review cases for which appeals were not filed. It is therefore possible that litigants appeal decisions written by black and white judges at different rates and based on different (perhaps race-based) criteria, a practice that could bias the results seen here.

Two considerations weigh against this sort of "appeals bias" story. First, the literature on state-level judicial evaluation suggests that attorneys have a lower professional opinion of black judges (Kourlis and Singer, 2008), a worldview that might lead them to appeal decisions authored by these judges at higher rates compared

	Whites	Blacks
N judges	1263	86
	(94%)	(6%)
N cases heard	10410	536
	(95%)	(5%)

Table 3.8: Number of U.S. district court judges compared to number of cases appealed (for cases 1960-2002).

to those written by white judges. If attorneys are indeed more likely to appeal decisions by black judges, then a higher proportion of black-authored cases will be frivolous or merit-less appeals. This mechanism would, if anything, skew bias the results downward: appeals judges should be more likely to uphold (rather than reverse) the opinions written by black judges compared to whites. Believing that there may be an appeals bias would mean that what we see here is actually a conservative estimate of the true effect.[19]

Second, the data appear to suggest, somewhat tentatively, that appeals are taken from black and white judges at comparable rates. In Table 3.8, I explicitly compare the number of black and white U.S. district court judges in the population compared to the number of cases written by these judges that are heard on appeal: their respective proportions are roughly similar, which indicates that appeals panels are hearing cases from white and black lower court judges at comparable rates. (Not only does this counsel against an "appeals bias" story, but it also provides support against the possibility that cases written by black and white judges are systematically settled out of

[19] To be consistent with the results seen here, this "appeals bias" would have to work differently: if anything, litigants would have to be appealing from white-authored cases at a higher rate. This would result in appealed opinions written by white lower-court judges being, on average, more frivolous. Thus, more of the white-authored opinions would be upheld (because they are more frivolous) and the black-authored opinions (not being appealed at a higher rate) would appear to be reversed more often in comparison. This is, however, the less likely of the two options: what limited literature exists in this area suggests that litigants are less satisfied with opinions authored by black judges, making them (if anything) more likely to appeal those (Kourlis and Singer, 2008).

court, or otherwise dismissed or published, at different rates.) It is worth noting, however, that I use here a random subset of cases (from 1960-2002), with few cases in each year and in each jurisdiction; because appeals rates vary across time and through circuits, having more than just a random subset of cases would be preferable.

Southern Effect. Another possibility is that the nature of the effect is particularly pronounced in, or driven by, particular jurisdictions – including possibly the jurisdictions in the South. Here, the implications could be ambiguous. The appeals courts in the South are acknowledged by legal observers to be among the most conservative (for example, the 4th and 5th Circuits), which would implicate the possibility that black judges in these jurisdictions are more liberal than white judges and that is why they are being overruled more. On the other hand, the Southern appeals courts could also have suffered from lingering effects of institutionalized segregation and openly racially charged attitudes, which would implicate the racial bias explanation behind the effects.

To check the possibility that the "black judges effect" varies across parts of the country, I fit a fixed-effects regression that controls for a wide variety of judge-specific characteristics while allowing the effect of black lower-court authorship to vary over circuits.[20] The predicted probabilities resulting from this analysis are presented in Figure 3.10. Even though it appears that the effect is particularly strong in the 5th, 8th, and 10th Circuits (headquartered in New Orleans, St. Louis, and Denver, respectively), it is actually impossible to statistically distinguish any appeals courts from each other; that is, under no model specification can we reject the null hypothesis that the "black

[20] In results not presented here, I also drop from the analysis all southern states. Dropping these observations increased the uncertainty around all estimates, and slightly decreased the magnitude of the "black judges effect," but the results did not lose statistical significance.

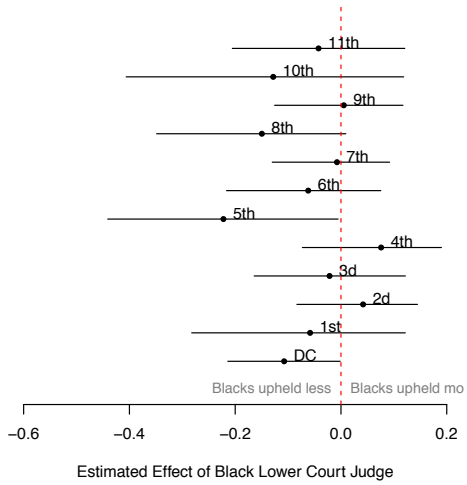

Figure 3.10: Effect of black authorship on probability case upheld, by U.S. Court of Appeals circuit. Predicted probabilities obtained via mixed-effect logit regression, controlling for (1) personal characteristics, (2) judicial experience and ABA ratings, and (3) case attributes. Judge-specific random effects included. Solid dots are point estimates; lines represent 95% confidence intervals.

judges" effect varies across any of the U.S. Courts of Appeals. Unfortunately, however, the lack of variance does little to possibly shed light on the causal mechanisms at play.

3.10 Conclusion

The results presented here suggest that discrepancies exist in how appeals courts have reviewed cases, with cases decided by black lower-court judges being more likely to be overturned than those authored by whites. This effect is robust and persists once we control for attributes that are possible proxies for judicial qualifications – e.g., quality of legal education, age, and professional experience, including private sector, academic, and prior judicial experience, as well as subjective measures like ABA quality ratings. Controls for the partisanship of the lower-court judge, as well as for the partisanship

and racial composition of the reviewing appeals panel, do not affect the results. This discrepancy between black and white judge is consistent across issue area and, with a few possible exceptions, across judicial circuits and regions of the country.

Regardless of the causal mechanisms, the implications are striking. Since John F. Kennedy, American Presidents have actively sought to appoint judges of color – not just African Americans, but also Hispanics, and Asian Americans – to the nation's highest courts. At the state and international level, too, efforts are underway to increase the proportion of judges who come from under-represented communities. The results presented here, however, call into question whether the mere appointment of these individuals is enough, and whether more ought to be done to ensure equality in the nation's courts. After all, if certain judges are being systematically overturned more often, then this raises questions about their relative long-term impact on the law, legal precedent, and the legal system – regardless of the reasons why.

To this extent, the results presented in this chapter actually represent just the tip of the iceberg in exploring the components of judicial evaluation and its relationship to descriptive representation – a topic previous unexplored in the judicial politics literature. Here, I touched upon just one singular ascriptive characteristic: the race (black or not black) of lower court judges. Whether a judge is African American is, however, just one facet of judicial identity, and we may think that the similar effects may exist for multiple racial groups (e.g., Asian Americans, Native Americans) and also for different ethnicities (Hispanics), religious groups (Jews, Catholics), and genders (women versus men) – not to mention multiple combinations of these identities (e.g., black women). In addition, if we think that heuristics or personal familiarity may play a role in how appeals panels reach decisions, then maybe we would find different rates of overturning between judges who attended the same law school or are otherwise

knowledgeable or friendly – that is, that a personal connection strengthens a bond that makes reversal less likely. Further research should help clarify the extent to which these and other attributes might play a role in appellate review.

Second, I examined only the effect of the author of the lower-court opinion. The identities of other actors – litigants, lawyers, government actors – are, however, crucially important to how judges decide cases. For example, the race of the litigator, which law firm he or she represents, and what law school he or she attended are all possible sources of influence, and the extent to which these attributes affect decision making is relatively unexplored outside the death penalty context. Similarly, we have a poor grasp on how these sorts of actors understand and evaluate judges. Studies on state judicial systems suggest that practicing attorneys have a skeptical view of minority and women judges, but unclear at this point is whether this potential bias has broader ripple effects through the rest of the judicial system – e.g., through increased or more aggressive appeals and higher rates of reversal.

Third, this is a study that relies on a quantitative analysis of a large subset of cases. Still remaining is a closer, qualitative look at the opinions authored by both black and white lower-court and appeals judges. Do black judges use different legal reasoning or articulate legal principles in a different way? Do black judges rely on different arguments in defining their opinions? Does the language used by appeals panels differ according to the identity of the legal actors involved? Given the results of this analysis, a qualitative examination into these issues would further shed light into why black judges are more likely to be overruled.

3.11 Appendix

American attitudes toward race and ethnicity have evolved rapidly over the last 50 years. More and more African Americans have entered into the ranks of the legal elite, serving successfully as judges, government lawyers, and litigators. This influx may have increased whites' exposure to black lawyers and therefore decreased potential racial bias. In addition, we may expect that the increased influx of African Americans into elite law schools (particularly those previously segregated) and into prominent legal practice also has decreased the reversal rate of African-American judges over time.

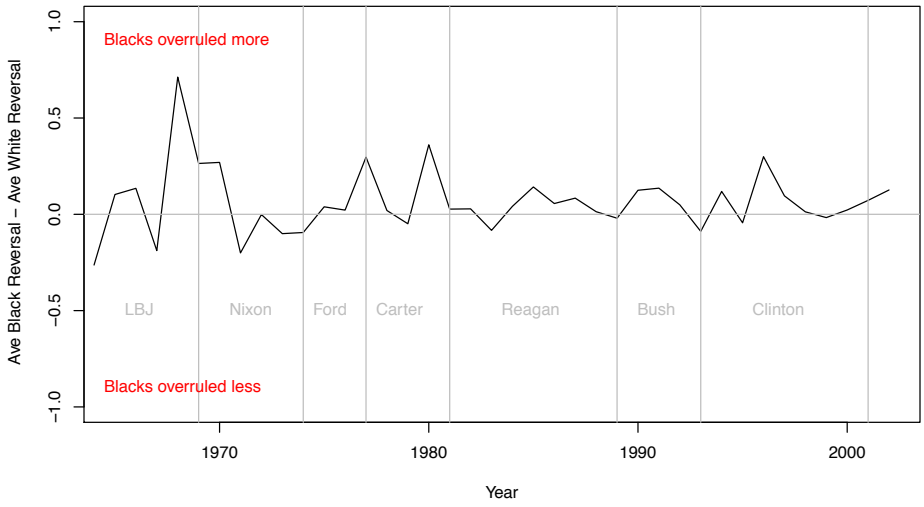

Figure 3.11: Difference in the average reversal rate of black U.S. District Court judges versus the average reversal rate of white U.S. District Court judges, by year.

However, the difference between black and white judges appears not to change over time (Figure 3.11). Indeed, once we take into account year effects in a fixed-effects regresssion (Figure 3.12), we're unable to rule out the null hypothesis that there is no

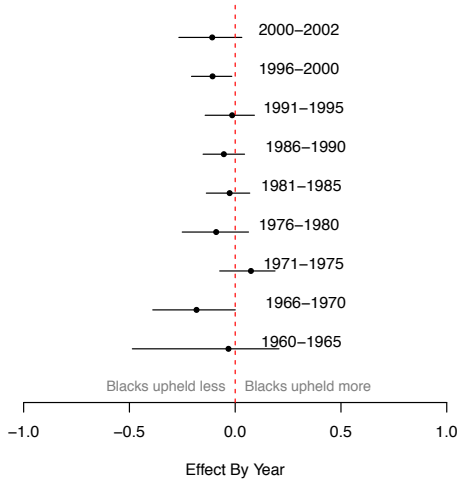

Figure 3.12: Effect of district judge race on the probability of being upheld at the Court of Appeals. Predicted probabilities obtained after fitting a fixed-effects logit regression, controlling for the same case and judicial characteristics as before. Judge-specific random effects included. Solid dots represent point estimates, while the lines represent 95% confidence intervals.

statistically significant fluctuation in the "black judges effect" across year cohorts. In other words, there is no evidence that suggests that the effect of having a black lower court judge varies over time or has attenuated across the years.

Admittedly, however, the data extend only through 2002. Given the changes brought about by the election of Barack Obama, more data would be necessary to answer the question definitively.

References

Adida, C.L, D.D Laitin and M.A Valfort. 2010. "Identifying barriers to Muslim integration in France." *Proceedings of the National Academy of Sciences* 107(52):1–7.

American Bar Association. 2009. "Standing Committee on the Federal Judiciary: What It is and How it Works.".

Angrist, J.D., G.W. Imbens and D.B. Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91(434).

Appiah, Kwame Anthony. 1986. The Uncompleted Argument: Du Bois and the Illusion of Race. In *"Race," Writing and Difference*, ed. Jr. Henry Louis Gates. Chicago: University of Chicago Press.

Asmussen, N. 2011. "Nontraditional Judicial Nominees: President's Delight and Senators' Dismay?" *Legislative Studies Quarterly* .

Bates, B., M. Maechler and B. Bolker. 2011. "lme4: Linear mixed-effects models using S4 classes.".

Bertrand, M. and S. Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94(4):991–1013.

Bobo, L and D Johnson. 2004. "A taste for punishment: Black and white Americans' views on the death penalty and the war drugs." *Du Bois Review: Social Science Research on Race* .

Boker, SM, JF Cohn, BJ Theobald and I Matthews. N.d. "Something in the way we move: Motion dynamics, not perceived sex, influence head movements in conversation." *... Experimental Psychology: ...* . Forthcoming.

Boyd, C.L. 2011. "Federal District Court Judge Ideology Data.".
URL: *http://cLboyd.net/ideology.html*

Boyd, C.L., L. Epstein and A.D. Martin. 2010. "Untangling the causal effects of sex on judging." *American Journal of Political Science* 54(2):389–411.

Brader, T., N.A. Valentino and E. Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American Journal of Political Science* 52(4):959–978.

Brudney, J.J., S. Schiavoni and D.J. Merritt. 1999. "Judicial Hostility toward Labor Unions–Applying the Social Background Model to a Celebrated Concern." *Ohio State Law Journal* 60:1675.

Chang, Jonathan, Itamar Rosenn, Lars Backstrom and Cameron Marlow. 2010. ePluribus: Ethnicity on Social Networks.

Cook, T.D., D.T. Campbell and A. Day. 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston.

Cosmides, L, John Tooby and Robert Kuzban. 2003. "Perceptions of race." *Trends in Cognitive Sciences* 7(4):173–179.

Cutler, D.M., R.G. Fryer and E.L. Glaeser. 2005. "Racial Differences in Life Expectancy: The Impact of Salt, Slavery, and Selection.".

de Rohan Barondes, R. 2009. "ABA Ratings of Federal District Court Judges and the Likelihood of a Shepard's Warning Signal.".

Deaux, Kay. 1985. "Sex and gender." *Annual Review of Psychology* 36:49–81.

Dehejia, R.H. and S. Wahba. 2002. "Propensity score-matching methods for nonexperimental causal studies." *Review of Economics and Statistics* 84(1):151–161.

Dolnick, Sam. 2011. "Ethnic Differences Emerge in Plastic Surgery." *New York Times* (2/18/2011).

Epstein, L., A.D. Martin, J.A. Segal and C. Westerland. 2007. "The judicial common space." *Journal of Law, Economics, and Organization* 23(2):303.

Epstein, L., J. Knight and A.D. Martin. 2003. "Norm of Prior Judicial Experience and Its Consequences for Career Diversity on the US Supreme Court, The." *Cal. L. Rev.* 91:903.

Faulkner, W. 1990. "Light in August. 1932." *London: Vintage* .

Fogel, R. 1994. "Economic growth, population theory, and physiology: The bearing of long-term processes on the making of economic policy." *nber.org* .

Gates, H.L. 1997. "The Passing of Anatole Broyard." *Thirteen Ways of Looking at a Black Man* pp. 180–214.

Gelman, A. and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge.

Giles, M.W., V.A. Hettinger and T. Peppers. 2001. "Picking federal judges: A note on policy and partisan selection agendas." *Political Research Quarterly* 54(3):623–641.

Ginther, D.K., W.T. Schaffer, J. Schnell, B. Masimore, F. Liu, L.L. Haak and R. Kington. 2011. "Race, ethnicity, and NIH research awards." *Science* 333(6045):1015–1019.

Gonzales, A. 2001. "Letter from Alberto Gonzales, U.S. Attorney General, to Martha Barnett, President of the American Bar Association.".

Gottschall, J. 1983. "Carter's Judicial Appointments: The Influence of Affirmative Action and Merit Selection on Voting on the U.S. Courts of Appeals." *Judicature* 67:165.

Greenwald, A.G, D.E McGhee and J.L.K Schwartz. 1998. "Measuring individual differences in implicit cognition: The implicit association test." *Journal of Personality and Social Psychology* 74(6):1464–1480.

Greiner, D.J. and D.B. Rubin. 2010. "Causal Effects of Perceived Immutable Characteristics." *Review of Economics and Statistics* .

Griffin, J.H. 1996. *Black Like Me*. NAL.

Guttentag, M. and P.F. Secord. 1983. *Too many women?: The sex ratio question*. Sage Publications.

Haire, S.B. 2001. "Rating the Ratings of the American Bar Association Standing Committee on Federal Judiciary." *Just. Sys. J.* 22:1.

Halsell, G. 1969. *Soul Sister*. World Pub. Co.

Hausman, D.M. 1998. *Causal asymmetries*. Cornell University Press.

Heckman, J.J. 1998. "Detecting discrimination." *The Journal of Economic Perspectives* 12(2):101–116.

Heckman, J.J. 2005. "Rejoinder: response to Sobel." *Sociological Methodology* 35(1):135.

Heckman, J.J. and P. Siegelman. 1993. "The urban Institute audit studies: Their methods and findings." *Clear and convincing evidence: Measurement of discrimination in America* pp. 187–258.

Ho, D.E. 2005. "Why Affirmative Action Does Not Cause Black Students to Fail the Bar." *Yale Law Journal* 114(8):1997–2005.

Ho, D.E., K. Imai, G. King and E.A. Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political Analysis* .

Hochschild, JL and V. Weaver. 2010. *Perspectives on Politics* 8(3).

Holland, P.W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–960.

Holland, P.W. 2003. "Causation and Race." *ETS Research Report* .

Huber, G.A. and J.S. Lapinski. 2006. "The "race card" revisited: Assessing racial priming in policy contests." *American Journal of Political Science* 50(2):421–440.

Hume, D. 2003. *A treatise of human nature*. Dover Pubns.

Humphreys, M. and J.M. Weinstein. 2008. "Who fights? The determinants of participation in civil war." *American Journal of Political Science* 52(2):436–455.

IAALS. 2008. "The Bench Speaks on Judicial Performance Evaluation: A Survey of Colorado Judges.".

Iacus, S., G. King and G. Porro. 2009. "CEM: Software for Coarsened Exact Matching.".

Iacus, S.M., G. King and G. Porro. 2011. "Multivariate Matching Methods That are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* .

Imai, K., G. King and E.A. Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the royal statistical society: series A (statistics in society)* 171(2):481–502.

Imbens, Guido W. and Donald B. Rubin. 2010. "Causal Inference in Statistics and Social Sciences.".

Just the Beginning Foundation. 2012. "Integration of the Federal Judiciary.". **URL:** *http://www.jtbf.org*

Kastellec, J.P. 2011. "Racial Diversity and Judicial Influence on Appellate Courts.".

Katz, Lawrence F, Jeffrey R Kling and Jeffrey B Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment*." *Quarterly Journal of Economics* (May):607–654.

Kearney, R.C. 1999. "Judicial Performance Evaluation in the States." *Public Administration Quarterly* 22:468–489.

Keele, L. 2010. "An Overview of rbounds: An R package for Rosenbaum Bounds Sensitivity Analysis with Matched Data." *White Paper. Columbus, OH* pp. 1–15.

Kim, C.J. and T. Lee. 2001. "Interracial politics: Asian Americans and other communities of color." *PS: Political Science and Politics* 34(03):631–637.

King, G. 1991. "" Truth" Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* pp. 1047–1053.

King, G. and L. Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131.

King, G., R.O. Keohane and S. Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton Univ Pr.

Kourlis, R.L. and J.M. Singer. 2008. "Performance Evaluation Program for the Federal Judiciary." *Denv. UL Rev.* 86:7.

Kurzban, Robert, J Tooby and L Cosmides. 2001. "Can race be erased? Coalitional computation and social categorization." *Proceedings of the National Academy of Sciences* 98(26):15387–15392.

Lindgren, J.T. 2001. "Examining the American Bar Association's Ratings of Nominees to the US Courts of Appeals for Political Bias, 1989-2000." *Northwestern Law Legal Working Paper Series* p. 37.

Locke, J. 1847. *An essay concerning human understanding*. Troutman & Hayes.

López, Ian F Haney. 1994. "The Social Construction of Race: Some Observations on Illusion, Fabrication, and Choice." *Harv C.R.-C.L. L. Rev.* 29(1):1–62.

Lott, J. 2001. "American Bar Association, Judicial Ratings, and Political Bias, The." *Journal of Law & Politics* 17:41.

Lott, J. 2006. "Pulling Rank." *The New York Times* .
URL: *http://www.nytimes.com/2006/01/25/opinion/25Lott.html*

Martin, E. and B. Pyle. 1999. "Gender, Race, and Partisanship on the Michigan Supreme Court." *Alb. L. Rev.* 63:1205.

Mendelberg, T. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton Univ Pr.

Menzies, P. and H. Price. 1993. "Causation as a secondary quality." *British Journal for the Philosophy of Science* 44(2):187–203.

Mill, S.J. 1884. "A system of logic.".

Miller, J.M. and J.A. Krosnick. 2000. "News media impact on the ingredients of presidential evaluations: Politically knowledgeable citizens are guided by a trusted source." *American Journal of Political Science* 44(2):301–315.

Mirengoff, P. 2010. "The American Bar Association Exposes Its Liberal Bias Once Again.".
URL: *http://www.powerlineblog.com/archives/2010/02/025696.php*

Navara, K.J. 2009. "Humans at tropical latitudes produce more females." *Biology letters* 5(4):524.

Neumark, D., R.J. Bank and K.D. Van Nort. 1996. "Sex discrimination in restaurant hiring: an audit study." *The Quarterly Journal of Economics* 111(3):915–941.

Nisbett, R.E. and D. Cohen. 1996. *Culture of honor: The psychology of violence in the South.* Westview Press.

Pager, D. 2007. "The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future." *The Annals of the American Academy of Political and Social Science* 609(1):104.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(March):937–75.

Pearl, J. 2000. *Causality: models, reasoning, and inference*. Cambridge Univ Pr.

Pelander, A.J. 1998. "Judicial Performance Review in Arizona: Goals, Practical Effects and Concerns." *Ariz. St. LJ* 30:643.

Peresie, J.L. 2005. "Female Judges Matter: Gender and Collegial Decisionmaking in the Federal Appellate Courts." *Yale Law Journal* 114(7):1759–1892.

Pinello, D.R. 2003. *Gay Rights and American law*. Cambridge Univ Pr.

Poole, K.T. 1998. "Recovering a basic space from a set of issue scales." *American Journal of Political Science* pp. 954–993.

Raben, R. 2011. "The ABA Ratings and Minority Nominees: Shedding Light on Disparate Impact.".
URL: *www.acslaw.org/acsblog/the-aba-ratings-and-minority-nominees-shedding-light-on-disparate-impact*

Rosenbaum, P.R. 2002. *Observational studies*. Springer Verlag.

Rubin, D.B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5):688–701.

Rubin, D.B. 1978. "Bayesian inference for causal effects: The role of randomization." *The Annals of Statistics* 6(1):34–58.

Rubin, D.B. 2005. "Causal inference using potential outcomes." *Journal of the American Statistical Association* 100(469):322–331.

Saks, M.J. and N. Vidmar. 2001. "Flawed Search for Bias in the American Bar Association's Ratings of Prospective Judicial Nominees: A Critique of the Lindgren Study." *Journal of Law & Politics* 17:219.

Sander, R.H. 2004. "A systemic analysis of affirmative action in American law schools." *Stanford Law Review* pp. 367–483.

Savage, C. 2011. "Ratings Shrink President's List for Judgeships.".

Scherer, N. 2004. "Blacks on the Bench." *Political Science Quarterly* pp. 655–675.

Schuyler, G.S. 1971. "Black No More. 1931." *New York: Collier* .

Segal, J.A. 2000. "Representative Decision Making on the Federal Bench: Clinton's District Court Appointees." *Political Research Quarterly* 53(1):137.

Segal, J.A. and H.J. Spaeth. 2002. *The Supreme Court and the attitudinal model revisited*. Cambridge Univ Pr.

Sekhon, J.S. 2009. "Opiates for the matches: Matching methods for causal inference." *Annual Review of Political Science* 12:487–508.

Sen, M. and O.T. Wasow. 2011. "Reconciling Race and Causation: Methods to Extract Meaningful Causal Inferences About Race.".
URL: *http://scholar.harvard.edu/msen/files/sen_wasow_causality.pdf*

Sisk, G.C., M. Heise and A.P. Morriss. 1998. "Charting the Influences on the Judicial Mind: An Empirical Study of Judicial Reasoning." *New York University Law Review* 73(5):1377–500.

Smelcer, S.N., A. Steigerwalt and R.L. Vining Jr. 2011. "Bias and the Bar: Evaluating the ABA Ratings of Federal Judicial Nominees." *Political Research Quarterly* .

Smith Jr, F.O. 2005. "Gendered Justice: Do Male and Female Judges Rule Differently on Questions of Gay Rights?" *Stanford Law Review* pp. 2087–2134.

Sniderman, P.M. and T.L. Piazza. 1993. *The Scar of Race*. Belknap Press.

Songer, Donald, Ahslyn K. Kuersten and Susan B. Haire. 2007. "The United States Courts of Appeals Database.".

Spohn, C. 1990. "The Sentencing Decisions of Black and White Judges: Expected and Unexpected Similarities." *Law & Society Review* 24(5):1197–1216.

Steele, C. 1997. "A threat in the air: How stereotypes shape intellectual identity and performance." *American psychologist* 52(6):613–629.

Sunstein, C.R., D. Schkade, L.M. Ellman and A. Sawicki. 2006. *Are Judges Political?* Brookings Institution Press.

Survey, Synovate AsiaBUS. 2004. "Skin lightening products in Asia - a bright future." *In:fact* June.
URL: *http://www.synovate.com/consumer-insights/infact/issues/200406/*

Tetreault, S. 2010. "Reid Criticizes Lawyers Group." *Law Vegas Review-Journal* .

Thomas, Katie and Brett Zarda. 2011. "In N.C.A.A., Question of Bias Over a Sickle-Cell Test." *New York Times* April(11).

Valentino, N.A., V.L. Hutchings and I.K. White. 2002. "Cues that matter: How political ads prime racial attitudes during campaigns." *American Political Science Review* 96(01):75–90.

Vining, R.L., A. Steigerwalt and S.N. Smelcher. 2009. "Bias and the Bar: Evaluating the ABA Ratings of Federal Judicial Nominees." *Unpublished draft available on ssrn. com (http://papers. ssrn. com/sol3/papers. cfm* .

Von Wright, G.H. 1971. *Explanation and understanding.* Cornell University Press.

Walker, T.G. and D.J. Barrow. 2009. "The diversification of the federal bench: Policy and process ramifications." *The Journal of Politics* 47(02):596–617.

Walton, G. M and G. L Cohen. 2011. "A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students." *Science* 331(6023):1447–1451.

Welch, S., M. Combs and J. Gruhl. 1988. "Do Black Judges Make a Difference?" *American Journal of Political Science* 32(1):126–136.

West, Candace and Don H Zimmerman. 1987. "Doing gender." *Gender & society* 1(2):125–151.

Whelan, E. 2010. "Re: The ABA and Ninth Circuit Nominee Goodwin Liu." *The National Review Online* .
URL: *http://www.nationalreview.com/bench-memos/49265/re-aba-and-ninth-circuit-nominee-goodwin-liu/ed-whelan*

White, I. 2007. "When race matters and when it doesn't: Racial group differences in response to racial cues." *American Political Science Review* 101(2):339–354.

Wood, R., S.R. Lazos and M. Waters. 2010. "Sacrificing Diversity for'Quality': How Judicial Performance Evaluations are Failing Women and Minorities." *Scholarly Commons at UNLV Law* .

Yinger, J. 1986. "Measuring racial discrimination with fair housing audits: Caught in the act." *The American Economic Review* 76(5):881–893.

Zuk, Gary, Deborah J. Barrow and Gerard Gryski. 2009. "Multi-User Database on the Attributes of United States Distric Court Judges, 1801-2000.". ICPSR 4553, Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.